Estimation of Regression Function Using Shannon's Entropy*

Millie Yi Mao[†] Aman Ullah[‡]

Abstract

We introduce an information theoretic approach—Shannon's entropy to specify econometric functions as an alternative to avoid parametric assumptions. We investigate the performances of Shannon's entropy in estimating the regression (conditional mean) and response (derivative) functions. We have demonstrated that they are easy to implement, and are advantageous over parametric models and nonparametric kernel techniques.

Key Words: Shannon's entropy, maximum entropy distributions, regression function, response function.

1. Introduction

In the literature of estimation, specification, and testing of econometric models, many parametric assumptions have been made. For example, a regression function is often considered to be linear. However, parametric econometrics has drawbacks since particular specifications may not capture the true data generating process. As a matter of fact, the true functional forms of econometric models are hardly known. Misspecification of parametric econometric models may therefore result in invalid conclusions and implications. Alternatively, data-based econometric methods can be adopted to avoid the disadvantages of parametric econometrics and implemented into practice. One widely-used approach is the nonparametric kernel technique, see Ullah (1988), Pagan and Ullah (1999), Li and Racine (2007) and Henderson and Parmeter (2015). However, nonparametric kernel procedures have some deficiencies, such as the "curse of dimensionality" and a lack of efficiency due to a slower rate of convergence of the variance to zero. In view of this, we propose a new information theoretic (IT) procedure for econometric model specification by using classical maximum entropy formulation. This is consistent, efficient, and based on minimal distributional assumptions.

Shannon (1948) derived the entropy (information) measure. Using Shannon's entropy measure Jaynes (1957a, 1957b) developed the maximum entropy principle to infer probability distribution. Entropy is a measure of a variable's average information content, and its maximization subject to some moments and normalization provides a probability distribution of the variable. The resulting distribution is known as the maximum entropy distribution; see more on this in Zellner and Highfield (1988), Ryu (1993), Golan et al. (1996), Harte et al. (2008), Judge and Mittelhammer (2011) and Golan (2018). We note that the joint probability distribution based on the maximum entropy approach is a purely data-driven distribution where parametric assumptions are avoided, and this distribution can be used to determine the regression function (conditional mean) and its response function (derivative function) which are of interest to empirical researchers.

We organize this paper in the following order. We present the IT based regression and response functions using a bivariate maximum entropy distribution in Section 2. A recur-

*We thank the seminar participants at conference organized by Info-Metrics Institute, American University in November 2016, in its economics department in September 2017, and in JSM (Vancouver) July-August 2018.

[†]Department of Economics, University of California, Riverside, CA 92521. E-mail: ymao004@ucr.edu

[‡]Department of Economics, University of California, Riverside, CA 92521. E-mail: aman.ullah@ucr.edu

sive integration process is developed for their implementations. In Section 3 we carry out a simulation example to illustrate the small sample efficiency of our methods. In Section 4, we present asymptotic theory on our IT based regression and response function estimators. In Section 5, we draw conclusions.

2. Estimation of Distribution, Regression, and Response Functions

We consider $\{y_i, x_i\}$, $i = 1, \ldots, n$ independent and identically distributed observations from an absolutely continuous bivariate distribution $f(y, x)$. Suppose the conditional mean of y given x exists and it provides a formulation for the regression model as

$$\begin{aligned} y &= E(y|x) + u \\ &= m(x) + u, \end{aligned} \tag{1}$$

where the error term u is such that $E(u|x) = 0$, and the regression function (conditional mean) is

$$E(y|x) = m(x) = \int_y y \frac{f(y, x)}{f(x)} dy. \tag{2}$$

When the joint distribution of y and x is not known, which is often the case, we propose the IT based maximum entropy method to estimate the densities of the random variables and introduce a recursive integration method to solve the conditional mean of y given x.

2.1 Maximum Entropy Distribution Estimation: Bivariate and Marginal

Suppose x is a scalar and the marginal density of it is unknown. Our objective is to approximate the marginal density $f(x)$ by maximizing the information measure (Shannon's entropy) subject to some constraints. That is

$$\operatorname*{Max}_f H(f) = -\int_x f(x) \log f(x) dx,$$

subject to

$$\int_x \phi_m(x) f(x) dx = \mu_m = E\phi_m(x), \, m = 0, 1, ..., M,$$

where $\phi_m(x)$ are known functions of x. $\phi_0(x) = \mu_0 = 1$. See, for example, Jaynes (1957a, 1957b) and Golan (2018). The total number of constraints is $M + 1$. In particular, $\phi_m(x)$ can be moment functions of x. We construct the Lagrangian

$$\mathcal{L}(\lambda_0, \lambda_1, \ldots, \lambda_M) = -\int_x f(x) \log f(x) dx + \sum\nolimits_{m=0}^{M} \lambda_m \left(\mu_m - \int_x \phi_m(x) f(x) dx \right),$$

where $\lambda_0, \lambda_1, \ldots, \lambda_M$ represent Lagrange multipliers. The solution has the form

$$\begin{aligned} f(x) &= \exp\left[-\sum\nolimits_{m=0}^{M} \lambda_m \phi_m(x) \right] = \frac{\exp\left[-\sum_{m=1}^{M} \lambda_m \phi_m(x) \right]}{\int_x \exp\left[-\sum_{m=1}^{M} \lambda_m \phi_m(x) \right] dx} \\ &\equiv \frac{\exp\left[-\sum_{m=1}^{M} \lambda_m \phi_m(x) \right]}{\Omega(\lambda_m)}, \end{aligned}$$

where λ_m is the Lagrange multiplier corresponding to constraint $\int_x \phi_m(x) f(x) dx = \mu_m$, and λ_0 (with $m = 0$) is the multiplier associated with the normalization constraint. With some simple algebra, it can be easily shown that $\lambda_0 = \log \Omega\left(\lambda_m\right)$ is a function of other multipliers. Replacing $f\left(x\right)$ and λ_0 into $\mathcal{L}\left(\lambda_0, \lambda_1, \ldots, \lambda_M\right) = \mathcal{L}\left(\boldsymbol{\lambda}\right)$, we get

$$\mathcal{L}\left(\boldsymbol{\lambda}\right) = \sum\nolimits_{m=1}^{M} \lambda_m E\phi_m(x) + \lambda_0.$$

The Lagrange multipliers are solved by maximizing $\mathcal{L}\left(\boldsymbol{\lambda}\right)$ with respect to λ_m's. The above inferred density is based on minimal information and assumptions. It is the flattest density according to the constraints. In this case, the Lagrange multipliers are not only the inferred parameters characterizing the density function, but also capture the amount of information conveyed in each one of the constraints relative to rest of the constraints used. They measure strength of the constraints.

In particular, when $M = 0$, $f\left(x\right)$ is a constant and hence x follows a uniform distribution. When the first moment of x is known, $f\left(x\right)$ has the form of an exponential distribution. When the first two moments of x are known, $f\left(x\right)$ has the form of a normal distribution. Furthermore, if more moment information is given, i.e. $M \geq 3$, to estimate the Lagrange multipliers, we use the Newton method considered in the literature. See Mead and Papanicolaou (1984) and Wu (2003).

In the bivariate case, the joint density of y and x is obtained from maximizing the information criterion $H\left(f\right)$ subject to some constraints. Here, we assume the moment conditions up to 4th order are known. Then

$$\underset{f}{\text{Max}}\ H(f) = -\int_x \int_y f(y,x) \log f(y,x) dydx \tag{3}$$

subject to

$$\int_x \int_y y^{m_1} x^{m_2} f(y,x) dydx = \mu_{m_1 m_2} = E\left(y^{m_1} x^{m_2}\right),\ 0 \leq m_1 + m_2 \leq 4. \tag{4}$$

We construct the Lagrangian

$$\begin{aligned} \mathcal{L}\left(\boldsymbol{\lambda}, \lambda_{00}\right) \quad = \quad & -\int_x \int_y f(y,x) \log f(y,x) dydx \\ & + \sum\nolimits_{m_1=0}^{4} \sum\nolimits_{m_2=0}^{4} \lambda_{m_1 m_2} \left(\mu_{m_1 m_2} - \int_x \int_y y^{m_1} x^{m_2} f(y,x) dydx\right), \end{aligned} \tag{5}$$

where $\boldsymbol{\lambda} = \left(\lambda_{m_1 m_2}\right)_{14 \times 1}$ for all $1 \leq m_1 + m_2 \leq 4$. The solution of the joint density distribution yields the form

$$\begin{aligned} f\left(y,x\right) \quad &= \quad \exp\left[-\sum\nolimits_{m_1+m_2=0}^{4} \lambda_{m_1 m_2} y^{m_1} x^{m_2}\right] \\ &= \quad \frac{\exp\left[-\sum_{m_1+m_2=1}^{4} \lambda_{m_1 m_2} y^{m_1} x^{m_2}\right]}{\int_x \int_y \exp\left[-\sum_{m_1+m_2=1}^{4} \lambda_{m_1 m_2} y^{m_1} x^{m_2}\right] dydx} \\ &\equiv \quad \frac{\exp\left[-\sum_{m_1+m_2=1}^{4} \lambda_{m_1 m_2} y^{m_1} x^{m_2}\right]}{\Omega\left(\lambda_{m_1 m_2}\right)}, \end{aligned} \tag{6}$$

where $\lambda_{m_1 m_2}$ is the Lagrange multiplier that corresponds to the constraint $\int_x \int_y y^{m_1} x^{m_2} f(y,x) dydx = \mu_{m_1 m_2}$, and $\lambda_{00} = \log \Omega\left(\lambda_{m_1 m_2}\right)$ (with $m_1 + m_2 = 0$) is the multiplier associated with the normalization constraint which is a function of other multipliers. See, e.g., Golan (1988, 2018) and Ryu (1993).

For deriving our results in Section 2, we rearrange the terms in $f(y,x)$ and write

$$f(y,x) = \exp\left[-\left(\lambda_{04}x^4 + \lambda_{03}x^3 + \lambda_{02}x^2 + \lambda_{01}x + \lambda_{00}\right)\right] \times \exp\left\{-\left[\lambda_{40}y^4 + \lambda_{30}(x)y^3 + \lambda_{20}(x)y^2 + \lambda_{10}(x)\right]\right\} y \tag{7}$$

where

$$\lambda_{30}(x) = \lambda_{30} + \lambda_{31}x,\ \lambda_{20}(x) = \lambda_{20} + \lambda_{21}x + \lambda_{22}x^2,$$
$$\lambda_{10}(x) = \lambda_{10} + \lambda_{11}x + \lambda_{12}x^2 + \lambda_{13}x^3.$$

Replacing $f(y,x)$ and λ_{00} into $\mathcal{L}(\boldsymbol{\lambda}, \lambda_{00}) = \mathcal{L}(\boldsymbol{\lambda})$, we obtain the Lagrange multipliers by maximizing

$$\mathcal{L}(\boldsymbol{\lambda}) = \sum\nolimits_{m_1+m_2=1}^{4} \lambda_{m_1m_2}\mu_{m_1m_2} + \lambda_{00}. \tag{8}$$

The marginal density of x is computed by integrating $f(y,x)$ over the support of y,

$$\begin{aligned} f(x) &= \int_y f(y,x)dy \\ &= \exp\left[-\left(\lambda_{04}x^4 + \lambda_{03}x^3 + \lambda_{02}x^2 + \lambda_{01}x + \lambda_{00}\right)\right] \\ &\times \int_y \exp\left\{-\left[\lambda_{40}y^4 + \lambda_{30}(x)y^3 + \lambda_{20}(x)y^2 + \lambda_{10}(x)\right]\right\} y. \end{aligned} \tag{9}$$

We note that $f(x) = f(x, \boldsymbol{\lambda})$ and $f(y,x) = f(y,x,\boldsymbol{\lambda})$. When the Lagrange multipliers $\boldsymbol{\lambda}$ are estimated as $\hat{\boldsymbol{\lambda}}$ from (8), we get $\hat{f}(x) = f(x, \hat{\boldsymbol{\lambda}})$ and $\hat{f}(y,x) = f(y,x,\hat{\boldsymbol{\lambda}})$.

Although the above results are written under fourth order moment conditions in (4), they can be easily written when $0 \leq m_1 + m_2 \leq M$. We have considered fourth order moment conditions without any loss of generality since they capture data information on skewness and kurtosis.

2.2 Regression and Response Functions

Based on the bivariate maximum entropy joint distribution (7) and the marginal density (9), the conditional mean (regression function) of y given x is represented as

$$\begin{aligned} m(x) = E(y \mid x) &= \int_y y \frac{f(y,x)}{f(x)} dy \\ &= \frac{\int_y y \exp\left\{-\left[\lambda_{40}y^4 + \lambda_{30}(x)y^3 + \lambda_{20}(x)y^2 + \lambda_{10}(x)y\right]\right\} dy}{\int_y \exp\left\{-\left[\lambda_{40}y^4 + \lambda_{30}(x)y^3 + \lambda_{20}(x)y^2 + \lambda_{10}(x)y\right]\right\} dy}. \end{aligned} \tag{10}$$

Given the values of the Lagrange multipliers, we define

$$F_r(x) \equiv \int_y y^r \exp\left\{-\left[\lambda_{40}y^4 + \lambda_{30}(x)y^3 + \lambda_{20}(x)y^2 + \lambda_{10}(x)y\right]\right\} dy. \tag{11}$$

where $r = 0, 1, 2, \ldots$. The regression function $m(x)$ thus takes the form

$$m(x) = m(x, \boldsymbol{\lambda}^*) = \frac{F_1(x)}{F_0(x)} = \frac{F_1(x, \boldsymbol{\lambda}^*)}{F_0(x, \boldsymbol{\lambda}^*)}, \tag{12}$$

where $\boldsymbol{\lambda}^* = (\lambda_{m_1m_2})_{10\times 1}$ for all $1 \leq m_1 + m_2 \leq 4$ except λ_{0m_2} for $m_2 = 1, \ldots, 4$. When the Lagrange multipliers are estimated from (8) by Newton method,

$$\hat{m}(x) = m\left(x, \hat{\boldsymbol{\lambda}}^*\right) = \frac{F_1\left(x, \hat{\boldsymbol{\lambda}}^*\right)}{F_0\left(x, \hat{\boldsymbol{\lambda}}^*\right)}. \tag{13}$$

This is the IT nonparametric regression function estimator. Furthermore, the response function $\beta(x) = \frac{dm(x)}{dx}$ (derivative) can be written as

$$\beta(x) = \beta(x, \boldsymbol{\lambda}^*) = \frac{F_1'(x, \boldsymbol{\lambda}^*) F_0(x, \boldsymbol{\lambda}^*) - F_1(x, \boldsymbol{\lambda}^*) F_0'(x, \boldsymbol{\lambda}^*)}{F_0^2(x, \boldsymbol{\lambda}^*)}, \tag{14}$$

and its estimator is given by

$$\hat{\beta}(x) = \beta\left(x, \hat{\boldsymbol{\lambda}}^*\right) \tag{15}$$

We note that $F_r'(x)$ represents the first derivative of $F_r(x)$ with respect to x, $r = 0, 1, 2, \ldots$.

2.3 Recursive Integration

It is unlikely to solve out the exponential polynomial integrals in the numerator and denominator from (10) in explicit forms. Numerical methods can be used to solve the problem by integrating the exponential polynomial function at each value of x. However, for large sample size, numerical methods are quite computationally expensive and hence are not satisfactory. We develop a recursive integration method which can not only solve the conditional mean $m(x)$ but also reduce the computational cost significantly.

According to the definition of $F_r(x)$ in (11), the changes in F_0, F_1 and F_2 are given by

$$\begin{aligned} F_0' &= -\lambda_{30}'(x)F_3 - \lambda_{20}'(x)F_2 - \lambda_{10}'(x)F_1 \\ F_1' &= -\lambda_{30}'(x)F_4 - \lambda_{20}'(x)F_3 - \lambda_{10}'(x)F_2 \\ F_2' &= -\lambda_{30}'(x)F_5 - \lambda_{20}'(x)F_4 - \lambda_{10}'(x)F_3, \end{aligned} \tag{16}$$

where $\lambda'(x)$ denotes the first derivative of $\lambda(x)$ with respect to x. Due to the special properties of (11), integrals of higher order exponential polynomial functions can be represented by those of lower orders. Based on this fact, F_3, F_4 and F_5 in (16) are replaced by the linear combinations of F_0, F_1 and F_2, resulting in a system of linear equations

$$\begin{aligned} F_0'(x) &= \Lambda_{00}(x)F_0(x) + \Lambda_{01}(x)F_1(x) + \Lambda_{02}(x)F_2(x) \\ F_1'(x) &= \Lambda_{10}(x)F_0(x) + \Lambda_{11}(x)F_1(x) + \Lambda_{12}(x)F_2(x) \\ F_2'(x) &= \Lambda_{20}(x)F_0(x) + \Lambda_{21}(x)F_1(x) + \Lambda_{22}(x)F_2(x). \end{aligned} \tag{17}$$

Starting from an initial value x_0, for a very small increment h, we trace out $F_0(x)$, $F_1(x)$ and $F_2(x)$ over the entire range of x

$$\begin{aligned} F_0(x_0 + h) &\approx F_0(x_0) + F_0'(x_0)h \\ F_1(x_0 + h) &\approx F_1(x_0) + F_1'(x_0)h \\ F_2(x_0 + h) &\approx F_2(x_0) + F_2'(x_0)h \end{aligned} \tag{18}$$

The IT estimators $\hat{m}(x)$ in (13) and $\hat{\beta}(x)$ in (15) are thus evaluated using (17) and (18) with $\boldsymbol{\lambda}^*$ replaced by $\hat{\boldsymbol{\lambda}}^*$. The results for finite domain integration are similar to the above.

3. Simulation Example

Here we consider a nonlinear data generating process (DGP) to evaluate the performance of our proposed IT estimator of regression and response functions.

The true model considered is a nonlinear function[1]

$$y_i = -\frac{1}{5}\log\left(e^{-2.5} + 2e^{-5x_i}\right) + u_i \tag{19}$$

[1]This simulation example is similar to Rilstone and Ullah (1989).

where $i = 1, 2, \ldots, n$, the variables y_i and x_i are in $\log$ values, and x_i are independent and identically drawn from uniform distribution with mean 0.5 and variance $\frac{1}{12}$. The error term u_i follows independent and identical normal distribution with mean 0 and variance 0.01.

The goal is to estimate the response coefficient $\beta(x) = \frac{\partial y}{\partial x}$. Two parametric approximations considered are

$$\begin{aligned} Linear &: \; y_i = \beta_0 + \beta_1 x_i + u_i \\ Quadratic &: \; y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i. \end{aligned}$$

These two parametric models are not correctly specified. Thus, one can expect that the estimation of the response coefficients may be biased. Besides these two parametric models, local constant nonparametric estimation of the response coefficient is also of our interest as a comparison with our IT method estimator. The local constant (Nadaraya-Watson) nonparametric kernel estimator is $\tilde{m}(x) = \sum y_i w_i(x)$, where $w_i(x) = \frac{K((x_i - x)/b)}{\sum K((x_i - x)/b)}$ in which $K(\cdot)$ is a kernel function and b is the bandwidth, for example, see Pagan and Ullah (1999). We have used normal kernel and cross-validated bandwidth. The bias and root mean square error (RMSE) results from linear function, quadratic approximation, local constant nonparametric method and IT method are reported in Table 1, averaged over 1000 replications of sample size 200. The values of the response coefficients shown are evaluated at the population mean of x, which is 0.5. Standard errors are given in the parentheses. True value of the response coefficient $\beta(x = 0.5) = 0.6667$.

Table 1

	Linear	Quadratic	Nonparametric	IT
$\beta(x) = \frac{\partial y}{\partial x}$	0.6288	0.6296	0.6468	0.6550
	(0.0276)	(0.0263)	(0.0904)	(0.0268)
Bias	0.0379	0.0371	0.0199	**0.0117**
RMSE	0.0469	0.0455	0.0926	**0.0292**

The biases for nonparametric kernel and IT estimators are smaller than those under linear and quadratic approximations. However, nonparametric estimation yields a larger RMSE compared with the three other methods. Even though nonparametric and IT estimations both have the advantage of avoiding the difficulties associated with the functional forms, results have indicated that the IT method outperforms the nonparametric method. This may be because the rate of convergence for MSE to zero for the IT estimator is n^{-1} whereas that of nonparametric kernel estimator is known to be $(nb)^{-1}$ where b is small (Li and Racine (2007)).

4. Asymptotic Properties of IT Estimators

First, we define

$$\begin{aligned} \underset{14\times1}{\mathbf{Z}_i} &= \left(y_i, x_i, y_i^2, x_i^2, y_i^3, x_i^3, y_i^4, x_i^4, y_i x_i, y_i x_i^2, y_i^2 x_i, y_i x_i^3, y_i^3 x_i, y_i^2 x_i^2\right)^T, \qquad (20) \\ \underset{14\times1}{\hat{\boldsymbol{\mu}}} &= \left(\hat{\mu}_{10}, \hat{\mu}_{01}, \hat{\mu}_{20}, \hat{\mu}_{02}, \hat{\mu}_{30}, \hat{\mu}_{03}, \hat{\mu}_{40}, \hat{\mu}_{04}, \hat{\mu}_{11}, \hat{\mu}_{12}, \hat{\mu}_{21}, \hat{\mu}_{13}, \hat{\mu}_{31}, \hat{\mu}_{22}\right)^T, \\ \underset{14\times1}{\boldsymbol{\mu}} &= \left(\mu_{10}, \mu_{01}, \mu_{20}, \mu_{02}, \mu_{30}, \mu_{03}, \mu_{40}, \mu_{04}, \mu_{11}, \mu_{12}, \mu_{21}, \mu_{13}, \mu_{31}, \mu_{22}\right)^T, \end{aligned}$$

where $\hat{\mu}_{m_1 m_2} = \frac{1}{n} \sum_{i=1}^{n} y_i^{m_1} x_i^{m_2}$, $\mu_{m_1 m_2} = E\left(y_i^{m_1} x_i^{m_2}\right)$, $m_1, m_2 = 0, 1, 2, 3, 4$ and $1 \leq m_1 + m_2 \leq 4$, and all the bold letters represent vectors. Suppose the following assumptions hold.

1. $\mathbf{Z}_i$, $i = 1, \ldots, n$ are independent and identically distributed from $(\boldsymbol{\mu}, \Sigma)$.

2. $\Sigma = \text{COV}(\mathbf{Z}_i)$ is assumed to be positive semi-definite, where the diagonals of Σ are

$$\text{Var}\left(y_i^{m_1} x_i^{m_2}\right) = \mu_{(2m_1)(2m_2)} - \mu_{m_1 m_2}^2,$$

and the off-diagonals of Σ are

$$\text{Cov}\left(y_i^{m_1} x_i^{m_2}, y_i^{m_1^*} x_i^{m_2^*}\right) = \mu_{\left(m_1+m_1^*\right)\left(m_2+m_2^*\right)} - \mu_{m_1 m_2}\mu_{m_1^* m_2^*}.$$

3. $\mu_{\left(m_1+m_1^*\right)\left(m_2+m_2^*\right)} < \infty$, $\forall m_1, m_2, m_1^*, m_2^* = 0, 1, 2, 3, 4,\ m_1 + m_2 \leq 4, m_1^* + m_2^* \leq 4$.

Now we present the following proposition.

Proposition 1. Under assumptions 1 to 3, as n goes to ∞,

$$\sqrt{n}\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) \sim N\left(\mathbf{0}, \Sigma\right). \tag{21}$$

Now, suppose the unique solution for each Lagrange multiplier exists. Then from (8), the vector $\boldsymbol{\lambda} = (\lambda_{10}, \lambda_{01}, \lambda_{20}, \lambda_{02}, \lambda_{30}, \lambda_{03}, \lambda_{40}, \lambda_{04}, \lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{13}, \lambda_{31}, \lambda_{22})^T$ can be expressed as a function of $\boldsymbol{\mu}$, i.e.

$$\boldsymbol{\lambda} = g\left(\boldsymbol{\mu}\right) \text{ and } \hat{\boldsymbol{\lambda}} = g\left(\hat{\boldsymbol{\mu}}\right). \tag{22}$$

Since from Proposition 1, $\sqrt{n}\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) \sim N\left(\mathbf{0}, \Sigma\right)$ as $n \to \infty$, it follows that

$$\sqrt{n}\left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\right) \sim N\left(\mathbf{0}, g^{(1)}\left(\boldsymbol{\mu}\right)\Sigma g^{(1)}\left(\boldsymbol{\mu}\right)^T\right) \text{ as } n \longrightarrow \infty, \tag{23}$$

where $g^{(1)}\left(\boldsymbol{\mu}\right) = \frac{\partial g(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}$ is the first derivative of $g\left(\boldsymbol{\mu}\right)$ with respect to $\boldsymbol{\mu}$.

Using the results in Proposition 1 and (23), $\sqrt{n}\left(\hat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}^*\right)$ follows $N\left(\mathbf{0}, g^{*(1)}\left(\boldsymbol{\mu}\right)\Sigma g^{*(1)}\left(\boldsymbol{\mu}\right)^T\right)$ as $n \longrightarrow \infty$, where $\boldsymbol{\lambda}^* = g^*\left(\boldsymbol{\mu}\right)$ and $g^{*(1)}\left(\boldsymbol{\mu}\right) = \frac{\partial g^*(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T}$. We get the following proposition for $\hat{m}\left(x\right)$ and $\hat{\beta}\left(x\right)$.

Proposition 2. Under assumptions 1 to 3 and (23), the asymptotic distributions of $\hat{m}\left(x\right) = m\left(x, \hat{\boldsymbol{\lambda}}^*\right)$ and $\hat{\beta}\left(x\right) = \beta\left(x, \hat{\boldsymbol{\lambda}}^*\right)$ are given as $n \to \infty$,

$$\sqrt{n}\left(m\left(x, \hat{\boldsymbol{\lambda}}^*\right) - m\left(x, \boldsymbol{\lambda}^*\right)\right) \sim N\left(\mathbf{0}, m^{(1)}\left(x, \boldsymbol{\lambda}^*\right) g^{*(1)}\left(\boldsymbol{\mu}\right)\Sigma g^{*(1)}\left(\boldsymbol{\mu}\right)^T m^{(1)}\left(x, \boldsymbol{\lambda}^*\right)^T\right), \tag{24}$$

where $m^{(1)}\left(x, \boldsymbol{\lambda}^*\right) = \frac{\partial m(x, \boldsymbol{\lambda}^*)}{\partial \boldsymbol{\lambda}^{*T}}$ is the first derivative of $m\left(x, \boldsymbol{\lambda}^*\right)$ with respect to $\boldsymbol{\lambda}^*$. And

$$\sqrt{n}\left(\beta\left(x, \hat{\boldsymbol{\lambda}}^*\right) - \beta\left(x, \boldsymbol{\lambda}^*\right)\right) \sim N\left(\mathbf{0}, \beta^{(1)}\left(x, \boldsymbol{\lambda}^*\right) g^{*(1)}\left(\boldsymbol{\mu}\right)\Sigma g^{*(1)}\left(\boldsymbol{\mu}\right)^T \beta^{(1)}\left(x, \boldsymbol{\lambda}^*\right)^T\right), \tag{25}$$

where $\beta^{(1)}\left(x, \boldsymbol{\lambda}^*\right) = \frac{\partial \beta(x, \boldsymbol{\lambda}^*)}{\partial \boldsymbol{\lambda}^{*T}}$ is the first derivative of $\beta\left(x, \boldsymbol{\lambda}^*\right)$ with respect to $\boldsymbol{\lambda}^*$.

Also, we note that the convergence rates of $m\left(x, \hat{\boldsymbol{\lambda}}^*\right)$ and $\beta\left(x, \hat{\boldsymbol{\lambda}}^*\right)$ are each $\sqrt{n}$.

5. Conclusions

In this paper, we estimate regression and response through Shannon's entropy. The advantages of using IT method over parametric specifications and nonparametric kernel approaches have been explained by the simulation example. It can be a useful tool for practitioners due to its simplicity and efficiency. Asymptotic properties are established. The IT based estimators are shown to be $\sqrt{n}$ consistent and normal. Thus, it has a faster rate of convergence compared to the nonparametric kernel procedures. We feel the IT approach for specifying regression and response functions considered here may open a new path to address specification and other related issues in econometrics with many applications.

REFERENCES

Golan, A. (1988). "A Discrete Stochastic Model of Economic Production and A Model of Fluctuations in Production–Theory and Empirical Evidence." Ph.D. thesis, University of California, Berkeley.

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

Golan, A. (2018). "Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information." Oxford University Press, New York.

Harte, J., Zillio, T., Conlisk, E. and Smith, A.B. (2008). "Maximum Entropy and the State-Variable Approach to Macroecology." *Ecology* 89, 2700-2711.

Henderson, D. and Parmeter, C. (2015). "Applied Nonparametric Econometrics." Cambridge University Press.

Jaynes, E.T. (1957a). "Information Theory and Statistical Mechanics." *Physical Review* 106, 620-630.

Jaynes, E.T. (1957b). "Information Theory and Statistical Mechanics II." *Physical Review* 108, 171-190.

Judge, G and Mittelhammer, R. (2011). "An Information Theoretic Approach to Econometrics." Cambridge University Press.

Li, Q. and Racine, J. (2007). "Nonparametric Econometrics: Theory and Practice." Princeton University Press.

Mead, L.R. and Papanicolaou, N. (1984). "Maximum Entropy in the Problem of Moments." *Journal of Mathematical Physics* 25, 2404-2417.

Pagan, A. and Ullah, A. (1999). "Nonparametric Econometrics." Cambridge University Press.

Rilstone, P. and Ullah, A. (1989). "Nonparametric Estimation of Response Coefficients." *Communications in Statistics-Theory and Methods* 18, 2615-2627.

Ryu, H. K. (1993). "Maximum Entropy Estimation of Density and Regression Functions." *Journal of Econometrics* 56, 397-440.

Shannon, C.E. (1948). "A Mathematical Theory of Communications." *The Bell System Technical Journal* 27, 379–423, 623–656.

Wu, X. (2003). "Calculation of Maximum Entropy Densities with Application to Income Distribution." *Journal of Econometrics* 115, 347-354.

Ullah, A. (1988). "Non-Parametric Estimation of Econometric Functionals." *The Canadian Journal of Economics* 21, 625-658.

Zellner, A. and Highfield, R.A. (1988). "Calculation of Maximum Entropy Distributions and Approximation of Marginalposterior Distributions." *Journal of Econometrics* 37, 195-209.