

Article

A Concentrated, Nonlinear Information-Theoretic Estimator for the Sample Selection Model

Amos Golan ^{1,*} and Henryk Gzyl ²

¹ Department of Economics and the Info-Metrics Institute, American University, Kreeger Hall 104, 4400 Massachusetts Ave., NW, Washington, DC 20016-8029, USA

² Centro de Finanzas, IESA, Caracas, Venezuela; E-Mail: henryk.gzyl@iesa.edu.ve

* Author to whom correspondence should be addressed; E-Mail: agolan@american.edu.

Received: 16 April 2010; in revised form: 3 June 2010 / Accepted: 11 June 2010 /

Published: 14 June 2010

Abstract: This paper develops a semi-parametric, Information-Theoretic method for estimating parameters for nonlinear data generated under a sample selection process. Considering the sample selection as a set of inequalities makes this model inherently nonlinear. This estimator (i) allows for a whole class of different priors, and (ii) is constructed as an unconstrained, concentrated model. This estimator is easy to apply and works well with small or complex data. We provide a number of explicit analytical examples for different priors' structures and an empirical example.

Keywords: concentrated model; inequalities; information; maximum entropy; priors; sample selection

1. Introduction and Basic Model

The sample selection problem appears often in empirical studies of labor supply, individuals' wages and other topics. For small sample sizes the existing parametric [1] and semi-parametric estimators [2–5] have difficulties. Recently, [6], henceforth GMP, developed a semi-parametric, Information-Theoretic (IT) estimator for the sample-selection problem that performs well when the sample is small. This estimator is based on the IT generalized maximum entropy (GME) approach of [7] and [8]. GMP used a large number of sampling experiments to investigate and compare the small-sample behavior of their estimator relative to other estimators. GMP concluded that their IT estimator is the most stable estimator while the likelihood estimators predicted better within the sample for large enough samples.

Their IT estimator outperformed the AP estimator in most cases and in all small samples. Another set of experiments within the nonlinear framework appear in [9].

GMP specified their IT-GME model with bounds on the parameters and with finite and discrete support. Though, their IT-GME estimator performs relatively well, it still has some of the basic shortcomings of that estimator. It has finite and bounded supports for both signal and noise, it is not flexible enough to incorporate infinitely large bounds and continuous support spaces, and it is constructed as a constrained optimization estimator. The objective here is to extend the estimator discussed in GMP in three directions. First, we allow unbounded support spaces for all parameters. Second, we accommodate for a whole class of (discrete and continuous) priors. Third, we construct our estimator as an unconstrained concentrated model.

1.1. The Basic Sample Selection Model

For simplicity, we follow a common labor model discussed in [10]. Suppose individual h ($h=1, \dots, N$) values staying (working) at home at wage y_{1h}^* and can earn y_{2h}^* in the marketplace. If $y_{2h}^* > y_{1h}^*$, the individual works in the marketplace, $y_{1h} = 1$, and we observe the market value, $y_{2h} = y_{2h}^*$. Otherwise, $y_{1h} = 0$ and $y_{2h} = 0$.

The individual's value at home or in the marketplace depends (linearly) on demographic characteristics ($\mathbf{x}$):

$$y_{1h}^* = \mathbf{x}_{1h}^t \beta_1 + \varepsilon_{1h} \quad (1)$$

$$y_{2h}^* = \mathbf{x}_{2h}^t \beta_2 + \varepsilon_{2h} \quad (2)$$

where $\mathbf{x}_{1h}$ and $\mathbf{x}_{2h}$ are K_1 and K_2 -dimensional vectors, β_1 and β_2 are K_1 and K_2 -dimensional vectors of unknowns and “ t ” stands for “transpose”. This model can be expressed as

$$y_{1h} = \begin{cases} 1 & \text{if } y_{2h}^* > y_{1h}^* \\ 0 & \text{if } y_{2h}^* \leq y_{1h}^* \end{cases} \quad (3)$$

$$y_{2h} = \begin{cases} \mathbf{x}_{2h}^t \beta_2 + \varepsilon_{2h} & \text{if } y_{2h}^* > y_{1h}^* \\ 0 & \text{if } y_{2h}^* \leq y_{1h}^* \end{cases} \quad (4)$$

Our objective is to estimate β_1 and β_2 . Typically the researcher is interested primarily in β_2 .

Unlike the more traditional models, GMP constructed their model as a solution to a constrained optimization problem such that the information represented by the set of censored equations (3)-(4) enters the estimation as inequalities:

$$\mathbf{x}_{2h}^t \beta_2 + \varepsilon_{2h} = y_{2h}, \quad \text{if } y_{1h} = 1 \quad (5)$$

$$\mathbf{x}_{2h}^t \beta_2 + \varepsilon_{2h} > \mathbf{x}_{1h}^t \beta_1 + \varepsilon_{1h}, \quad \text{if } y_{1h} = 1 \quad (6)$$

$$\mathbf{x}_{2h}^t \beta_2 + \varepsilon_{2h} \leq \mathbf{x}_{1h}^t \beta_1 + \varepsilon_{1h}, \quad \text{if } y_{1h} = 0 \quad (7)$$

In our formulation, we use inequalities as well to represent all available information in the set of censored equations.

2. The Information-Theoretic Estimator

Rewrite equations (1)-(2) as finding γ_1 and γ_2 in $y_1^* = A_1\gamma_1$ and $y_2^* = A_2\gamma_2$, where the dependent variable is censored and where $\gamma_1 = \begin{pmatrix} \beta_1 \\ \varepsilon_1 \end{pmatrix}$, $A_1 = [X_1 \ I]$, $\gamma_2 = \begin{pmatrix} \beta_2 \\ \varepsilon_2 \end{pmatrix}$ and $A_2 = [X_2 \ I]$. We formulate the censored model (5)-(7) in the following way.

Let the constraint sets $C_i = C_{i,s} \times C_{i,n}$ ($i=1,2$). For each i , $C_{i,s}$ is an auxiliary closed, convex set used to model the a-priori constraints on the β 's. Similarly, the closed convex set $C_{i,n}$ is part of the specification of the “physical” nature of the noise and contains all possible realizations of ε . We view the coordinates ζ_i in $C_{i,s}$ and ν_i in $C_{i,n}$ as values of random variables distributed according to some probability measure $dP_i(\xi_i) \equiv dP_i(\zeta_i, \nu_i)$ such that their expectations (E) are

$$\beta_i = E_{P_i}[\zeta_i] \quad \text{and} \quad \varepsilon_i = E_{P_i}[\nu_i] \quad (8)$$

We note that the qualifier “prior” assigned to the Q probability measures is not the traditional Bayesian view. Rather, the Q_s is just a mathematical construct to transform the estimation problem into a variational problem. The Q_n , however, could be viewed as the probability measure describing the statistical nature of the noise. The process of estimation of the noise involves a tilting of the prior measure.

Given some (any) prior measures $dQ_i(\xi_i) \equiv dQ_i(\zeta_i, \nu_i) = dQ_{i,s}(\zeta_i)dQ_{i,n}(\nu_i)$ we search for densities $\rho_i(\xi_i)$ such that $dP_i = \rho_i(\xi_i)dQ_i(\xi_i)$ satisfies the system (3)-(4). This yields the parameter estimator β^* and the estimated residuals ε^* .

Next, let $S(P_i, Q_i)$ denotes the differential entropy divergence measure between the priors, Q_i , and the post-data (posteriors) P_i . This is just the continuous version of the Kullback-Liebler information divergence measure, also known as relative entropy (see [11–13]). Since the data are naturally divided into observed and unobserved parts, we divide the data into two subsets: J and J^c of $\{1, 2, \dots, N\}$. Next, rewrite the data (3)-(4) $y_1^* = A_1\gamma_1$ and $y_2^* = A_2\gamma_2$ as

$$y_1^* = \begin{pmatrix} y_1 \\ \bar{y}_1 \end{pmatrix} = \begin{pmatrix} B_1 \\ \bar{B}_1 \end{pmatrix} E_{P_1}[\xi_1]; \quad \text{and} \quad y_2^* = \begin{pmatrix} y_2 \\ \bar{y}_2 \end{pmatrix} = \begin{pmatrix} B_2 \\ \bar{B}_2 \end{pmatrix} E_{P_2}[\xi_2] \quad (9)$$

where the matrices B_1 and B_2 correspond to the rows of the matrices A_i ($i=1, 2$) labeled by the indices for which observations are available. For the indices in J the values y_2 are observed and $y_{2h}^* > y_{1h}^*$, whereas for the values in J^c all we know is that $\bar{y}_{1h}^* > \bar{y}_{2h}^*$.

Our “Basic (Primal) Problem” is the solution to

$$\text{Sup}_{(P_1, P_2)} \{S(P_1, Q_1) + S(P_2, Q_2) \mid y_2 = B_2 E_{P_2}[\xi_2], B_2 E_{P_2}[\xi_2] > B_1 E_{P_1}[\xi_1]; \bar{B}_2 E_{P_2}[\xi_2] \leq \bar{B}_1 E_{P_1}[\xi_1]\} \quad (10)$$

where the inequalities between vectors are taken to be component wise.

Next, we formulate the problem as a concentrated (unconstrained) entropy problem. To do so, we view the basic primal problem as a two stage problem, call it an “*equivalent primal problem*.” In the equivalent model the first stage consists of the standard Generalized Entropy problem (the equality portion of the model) for which a dual can be easily formulated.

The Equivalent primal problem is a solution to the two stage optimization problem:

$$\begin{aligned} & \text{Sup}\{\text{Sup}\{S(P_1, Q_1) + S(P_2, Q_2) \mid y_2 = B_2 E_{P_2}[\xi_2], \eta_1 = B_1 E_{P_1}[\xi_1]; \\ & \bar{\eta}_2 = \bar{B}_2 E_{P_2}[\xi_2]; \bar{\eta}_1 = \bar{B}_1 E_{P_1}[\xi_1]\} \mid y_2 > \eta_1, \bar{\eta}_2 \leq \bar{\eta}_1\} \end{aligned} \quad (11)$$

Theorem 2.1. The *equivalent primal problem* (11) is equivalent to the following (dual) problem

$$\text{Inf}\{\ln Z(\lambda_1, -\bar{\lambda}, \lambda_2, \bar{\lambda}) + \langle \lambda_1 + \lambda_2, y_2 \rangle \mid \lambda_1 \in \mathbb{R}_+^{|J|}, \lambda_2 \in \mathbb{R}_+^{|J|} \text{ and } \bar{\lambda} \in \mathbb{R}_+^{N-|J|}\} \quad (12)$$

where $\langle a, b \rangle$ denotes the Euclidean scalar (inner) product of the vectors a and b ,

$$\begin{aligned} Z(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) &= \iint_{C_1 \times C_2} e^{-\langle \lambda_1, B_1 \xi_1 \rangle} e^{-\langle \bar{\lambda}_1, \bar{B}_1 \xi_1 \rangle} e^{-\langle \lambda_2, B_2 \xi_2 \rangle} e^{-\langle \bar{\lambda}_2, \bar{B}_2 \xi_2 \rangle} dQ_1(\xi_1) dQ_2(\xi_2) \\ &= \int_{C_1} e^{-\langle \lambda_1, B_1 \xi_1 \rangle} e^{-\langle \bar{\lambda}_1, \bar{B}_1 \xi_1 \rangle} dQ_1(\xi_1) \int_{C_2} e^{-\langle \lambda_2, B_2 \xi_2 \rangle} e^{-\langle \bar{\lambda}_2, \bar{B}_2 \xi_2 \rangle} dQ_2(\xi_2) = Z_1(\lambda_1, \bar{\lambda}_1) Z_2(\lambda_2, \bar{\lambda}_2) \end{aligned}$$

and $(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2)$ are the four sets of Lagrange multipliers associated with (11). To carry out the procedure specified in (12) first set $\bar{\lambda}_1 = -\bar{\lambda}_2 = -\bar{\lambda}$, and then carry out the minimization.

Proof: See Appendix.

To confirm the uniqueness of the solution to problem (12), observe that the function

$$\ell(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) = \ln Z(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) + \langle \lambda_1, \eta_1 \rangle + \langle \bar{\lambda}_1, \bar{\eta}_1 \rangle + \langle \lambda_2, y_2 \rangle + \langle \bar{\lambda}_2, \bar{\eta}_2 \rangle$$

is strictly convex on its domain $\Psi(Q) = \{\lambda = (\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) \mid Z(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) < \infty\}$, and if $\ell(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) \rightarrow \infty$ as $\lambda = (\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) \rightarrow \partial\Psi$, then problem (12) has a unique solution, where " ∂ " is the boundary of the set Ψ . This is always true in the cases we consider here. A simple example in which it does not hold is $l(\lambda) = e^\lambda + \lambda y$ with $\mathbb{R}$ as domain. This has no minimum for a positive y .

Solving (12) yields $\lambda_1^*, \bar{\lambda}^*, \lambda_2^*$, which in turn yields the optimal maximum entropy (posterior) density.

$$\rho^*(\xi_1, \xi_2) = \frac{e^{-\langle \lambda_1^*, B_1 \xi_1 \rangle - \langle \bar{\lambda}^*, \bar{B}_1 \xi_1 \rangle} e^{-\langle \lambda_2^*, B_2 \xi_2 \rangle + \langle \bar{\lambda}^*, \bar{B}_2 \xi_2 \rangle}}{Z(\lambda_1^*, \bar{\lambda}^*) Z(\lambda_2^*, -\bar{\lambda}^*)} = \rho_1^*(\xi_1) \rho_2^*(\xi_2)$$

This density is naturally factored into a product of the maximum entropy densities of the two sets of equations. Therefore, ξ_1 and ξ_2 are independent with respect to the reconstructed density $dP^*(\xi_1, \xi_2) = \rho_1^*(\xi_1) \rho_2^*(\xi_2) dQ_1(\xi_1) dQ_2(\xi_2)$, and with respect to the original priors. Once P^* is solved, we follow (8), or (9), to get

$$\begin{pmatrix} \beta_i^* \\ \varepsilon_i^* \end{pmatrix} = E_{P_i^*}[\xi_i^*] = \int_{C_i} \xi_i \rho_i^*(\xi_i) dQ_i(\xi_i); \quad i = 1, 2 \quad (13)$$

With that generic formulation, we show below three analytic examples that cover a wide range of possible priors and support spaces for β and ε .

3. Large Sample Properties

Denote by β_{iN}^* the estimator of the true β_i when the sample size is N . Throughout this section we add a subscript N to all quantities introduced in Section 2 to remind us that the size of the data set is N . We show that $\beta_{iN}^* \rightarrow \beta_i$ and $\sqrt{N}(\beta_{iN}^* - \beta_i) \rightarrow N(0, V_i)$ as $N \rightarrow \infty$ in some appropriate way. The proof is similar in logic to Proposition 3.2 in [14]. We assume:

Assumption 3.1. For every sample size N , the minimizers of (12) are all in the interior of their domains: $\lambda_1^* \in \text{int}(\mathbb{R}_+^{[I]})$ and $\bar{\lambda}_1^* \in \text{int}(\mathbb{R}_+^{N-[I]})$ where “int” stands for interior.

Assumption 3.2. Let $\frac{1}{N} X_i^T X_i = \frac{1}{N} (B_i^T B_i + \bar{B}_i^T \bar{B}_i) = \frac{J}{N} \left(\frac{1}{J} B_i^T B_i \right) + \frac{N-J}{N} \left(\frac{1}{N-J} \bar{B}_i^T \bar{B}_i \right)$. Then, assume there exists $\alpha \in (0, 1)$ such that (i) $N \rightarrow \infty$ and $J \rightarrow \infty$ such that $(J/N) \rightarrow \alpha$ and (ii) assume that there exists two matrices W_i^o and W_i^u such that $\frac{1}{J} B_i^T B_i \rightarrow W_i^o$ and $\frac{1}{N-J} \bar{B}_i^T \bar{B}_i \rightarrow W_i^u$.

Note that $\frac{1}{N} X_i^T X_i \rightarrow W_i = \alpha W_i^o + (1-\alpha) W_i^u$.

Proposition 3.1. (Convergence in distribution.) Under Assumptions 3.1 and 3.2

- a) $\beta_{iN}^* \xrightarrow{D} \beta_i$ as $N \rightarrow \infty$, for $i=1, 2$.
- b) $\sqrt{N} (\beta_{iN}^* - \beta_i) \xrightarrow{D} N(0, V_i)$ as $N \rightarrow \infty$

where $\xrightarrow{D}$ stands for convergence in distribution and $V_i = \Sigma_i W_i^{-1} \Sigma_i$, where Σ_i is the covariance matrix of ς_i with respect to $dQ_i(\varsigma_i, \nu_i)$.

The approximate finite sample variance is

$$\sigma_i^{*2} = \frac{1}{N - K_i} \epsilon_i^{*T} \epsilon_i^* \text{ for } i = 1, 2 \text{ and } \epsilon_i^* = E_{P_i^*}[\nu_i] \text{ as is shown in (8) or similarly in (12).}$$

4. Analytic Examples

We discuss three examples, corresponding to assuming that the β s are either unbounded (Normal), bounded below (Gamma) and bounded below and above (Bernoulli). Under the normal priors, the minimum described in (12) can be explicitly computed. In the other cases, a numerical computation is necessary.

4.1. Normal Priors

Let the constraint space be $C = C_s \times C_n = \mathbb{R}^K \times \mathbb{R}^N$. Using the traditional view and centering the support spaces at zero, the prior—a product of two normal distributions—is

$$dQ(\xi) = \frac{\exp(-\langle \xi, \Sigma^{-2} \xi \rangle / 2)}{(2\pi)^{(N+K)/2} (\det \Sigma^2)^{1/2}} d\xi. \text{ The covariance } \Sigma \text{ has two diagonal blocks: } K \times K \text{ and } N \times N.$$

Without loss of generality, we assume that these two matrices are $\sigma_s^2 I_K$ and $\sigma_n^2 I_N$ respectively. Our basic model holds for the general covariance structure $\Sigma = \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix}$.

Formulating these priors within our model yields

$$\ln Z_1(\lambda_1, \bar{\lambda}_1) = \frac{1}{2} \left\{ \langle \lambda_1, B_1 \Sigma_1^{-2} B_1^T \lambda_1 \rangle - 2 \langle \bar{\lambda}_1, \bar{B}_1 \Sigma_1^{-2} B_1^T \lambda_1 \rangle + \langle \bar{\lambda}_1, \bar{B}_1 \Sigma_1^{-2} \bar{B}_1^T \bar{\lambda}_1 \rangle \right\}$$

$$\ln Z_2(\lambda_2, -\bar{\lambda}_1) = \frac{1}{2} \left\{ \langle \lambda_2, B_2 \Sigma_2^2 B_2' \lambda_2 \rangle + 2 \langle \bar{\lambda}, \bar{B}_2 \Sigma_2^2 B_2' \lambda_2 \rangle + \langle \bar{\lambda}, \bar{B}_2 \Sigma_2^2 \bar{B}_2' \bar{\lambda} \rangle \right\}$$

where Σ_1^2 is a diagonal $(K + N) \times (K + N)$ matrix, the first block being a $K \times K$ matrix with entries equal to $\sigma_{1,s}^2$ and the second block is a $N \times N$ matrix with entries equal to $\sigma_{1,n}^2$. That is, the priors on signal and noise spaces are iid normal random variables. Thus, problem (12) consists of finding the minimum of

$$\begin{aligned} \ln Z(\lambda_1, \bar{\lambda}_1, \lambda_2, -\bar{\lambda}_1) + \langle \lambda_1 + \lambda_2, y_2 \rangle &= \langle \lambda_1 + \lambda_2, y_2 \rangle \\ &+ \frac{1}{2} \left\{ \langle \lambda_1, B_1 \Sigma_1^2 B_1' \lambda_1 \rangle - 2 \langle \bar{\lambda}, \bar{B}_1 \Sigma_1^2 B_1' \lambda_1 \rangle + \langle \bar{\lambda}, \bar{B}_1 \Sigma_1^2 \bar{B}_1' \bar{\lambda} \rangle \right\} \\ &+ \frac{1}{2} \left\{ \langle \lambda_2, B_2 \Sigma_2^2 B_2' \lambda_2 \rangle + 2 \langle \bar{\lambda}, \bar{B}_2 \Sigma_2^2 B_2' \lambda_2 \rangle + \langle \bar{\lambda}, \bar{B}_2 \Sigma_2^2 \bar{B}_2' \bar{\lambda} \rangle \right\} \end{aligned}$$

over the set described in (12).

To verify that the minimizer of (12) occurs in the interior of the constraint set, we look at the first order conditions

$$\begin{aligned} B_1 \Sigma_1^2 B_1' \lambda_1 - B_1 \Sigma_1^2 \bar{B}_1' \bar{\lambda} + y_2 &= 0 \\ B_2 \Sigma_2^2 B_2' \lambda_2 + B_2 \Sigma_2^2 \bar{B}_2' \bar{\lambda} + y_2 &= 0 \\ -\bar{B}_1 \Sigma_1^2 \bar{B}_1' \bar{\lambda} + \bar{B}_1 \Sigma_1^2 B_1' \lambda_1 - \bar{B}_2 \Sigma_2^2 \bar{B}_2' \bar{\lambda} - \bar{B}_2 \Sigma_2^2 B_2' \lambda_2 &= 0 \end{aligned} \quad (14)$$

A feasible solution to (12) may lie inside the domain of the constraints and provides a solution.

Once the system is solved for $\lambda_1^*, \bar{\lambda}_1^*, \lambda_2^*$, the estimated densities are

$$dP_i^*(\xi_i) = \frac{e^{-\|\Sigma_i^{-1} \xi_i + \Sigma_i h_i^*\|^2 / 2}}{(2\pi)^{(N+K_i)/2} (\det \Sigma_i^2)^{1/2}} d\xi_i$$

which, as expected, are normally distributed. Defining

$$h_i = B_i' \lambda_i + \bar{B}_i' \bar{\lambda}_i = A_i' \mu_i \equiv \begin{pmatrix} X_i' \\ I \end{pmatrix} \begin{pmatrix} \lambda_i^* \\ \bar{\lambda}_i^* \end{pmatrix}$$

(recall that $\bar{\lambda}_2^* = -\bar{\lambda}_1^* = -\bar{\lambda}$ due to the constraints) we use (13) to get

$$\begin{pmatrix} \beta_i^* \\ \varepsilon_i^* \end{pmatrix} = E_{P_i^*}[\xi_i^*] = -\Sigma_i^2 A_i' \mu_i = - \begin{pmatrix} \sigma_{i,s}^2 & 0 \\ 0 & \sigma_{i,n}^2 \end{pmatrix} A_i' \mu_i = - \begin{pmatrix} \sigma_{i,s}^2 X_i' \mu_i \\ \sigma_{i,n}^2 \mu_i \end{pmatrix}$$

for $i = 1, 2$ and where $\sigma_{i,s}^2$ and $\sigma_{i,n}^2$ are matrices.

4.2. Gamma Priors

Let β 's be bounded below by 0. This can be easily generalized by an appropriate shifting of the support of the distributions. To show the generality of our model, we let the prior on the noise be normal thereby showing that one can use different priors for the signal and the noise.

The signal and noise constraint spaces respectively are $C_s = \mathbb{R}_+^K = [0, \infty)^K$ and $C_n = \mathbb{R}^N$. The prior is

$$dQ(\xi) = \left(\prod_{j=1}^K \frac{a^b \zeta_j^{b-1} e^{-a\zeta_j}}{\Gamma(b)} \right) d\zeta_j \frac{e^{-\langle v, \Sigma_n^{-2} v \rangle / 2}}{(2\pi)^{N/2} (\det \Sigma_n^2)} dv$$

Before specifying the concentrated entropy function, we study the matrix A_1 defined as $A_1 = (X_1 \ I) = \begin{pmatrix} B_1 \\ \bar{B}_1 \end{pmatrix} = \begin{pmatrix} D_1 & I_1 \\ \bar{D}_1 & \bar{I}_1 \end{pmatrix}$. Note that $\begin{pmatrix} D_1 \\ \bar{D}_1 \end{pmatrix}$ splits X_1 and $\begin{pmatrix} I_1 \\ \bar{I}_1 \end{pmatrix}$ splits the $N \times N$ identity matrix to match the splitting of X . The concentrated entropy function is

$$\ln Z_1(\lambda_1, -\bar{\lambda}) = \sum_{j=1}^K b \ln \left(\frac{a}{a + (D_1^t \lambda_1 + \bar{D}_1^t \bar{\lambda})_j} \right) + \sigma_{1,n}^2 (\lambda_1^2 + \bar{\lambda}^2)$$

(Note that $D_1^t \lambda_1 + \bar{D}_1^t \bar{\lambda} = X_1^t \mu_1$ and $D_2^t \lambda_2 - \bar{D}_2^t \bar{\lambda} = X_2^t \mu_2$.) A similar expression exists for $\ln Z_2(\lambda_2, -\bar{\lambda})$.

The problem (12) consists of minimizing

$$\begin{aligned} & \ln Z(\lambda_1, -\bar{\lambda}, \lambda_2, \bar{\lambda}) + \langle \lambda_1 + \lambda_2, y_2 \rangle \\ &= \sum_{j=1}^{K_1} b \ln \left(\frac{a}{a + (D_1^t \lambda_1 + \bar{D}_1^t \bar{\lambda})_j} \right) + \sum_{j=1}^{K_2} b \ln \left(\frac{a}{a + (D_2^t \lambda_2 - \bar{D}_2^t \bar{\lambda})_j} \right) \\ &+ \sigma_{1,n}^2 (\lambda_1^2 + \bar{\lambda}^2) + \sigma_{2,n}^2 (\lambda_2^2 + \bar{\lambda}^2) + \langle \lambda_1 + \lambda_2, y_2 \rangle \end{aligned}$$

Once $\lambda_1^*, \bar{\lambda}^*, \lambda_2^*$ are found, the optimal density is

$$dP_i^*(\xi_i) = \left(\prod_{j=1}^{K_i} \frac{(a + (X_i^t \mu_i)_j)^b \zeta_j^{b-1} e^{-((X_i^t \mu_i)_j + a)\zeta_j}}{\Gamma(b)} d\zeta_j \right) \frac{e^{-\|v + \sigma_{i,n}^2 \mu_i^*\|^2 / 2\sigma_{i,n}^2}}{(2\pi\sigma_{i,n}^2)^{N_i/2}} dv$$

The estimated parameters are

$$(\beta_i^*)_j = E_{P_i^*}[(\zeta_i)_j] = \frac{b}{(a + (X_i^t \mu_i)_j)}; \quad \text{for } j=1, \dots, K_i \text{ and } i=1, 2$$

The realized residuals are

$$(\varepsilon_i^*)_l = E_{P_i^*}[(v_i)_l] = -\sigma_{i,n}^2 (\mu_i^*)_l; \quad \text{for } l=1, \dots, N_i \text{ and } i=1, 2$$

4.3. Bernoulli Priors

This example represents another extreme case where it is assumed that the β 's are bounded. For simplicity, assume that we know that all β 's lie in the interval $[a, b]$, which makes $C_s = [a, b]^{K_i}$ the choice for all of the constraints on the signal space. For the noise component, we follow the previous formulation of normal priors.

With this background, the prior measure used is

$$dQ(\xi) = \left(\prod_{j=1}^K \frac{1}{2} (\delta_a(d\zeta_j) + \delta_b(d\zeta_j)) \right) \frac{e^{-\langle v, \Sigma_n^{-2} v \rangle / 2}}{(2\pi)^{N/2} (\det \Sigma_n^2)} dv$$

The concentrated entropies are

$$\ln Z_i(\lambda_i, \bar{\lambda})_i = \sum_{j=1}^{K_i} \ln \frac{1}{2} (e^{-g_{i,j}a} + e^{-g_{i,j}b}) + \sigma_n^2 \|\mu_i\|^2, \quad i=1, 2$$

where $g_i = D_i^t \lambda_i + \bar{D}_i^t \bar{\lambda}_i$ and $\mu_i = (\lambda_i \bar{\lambda}_i)^t$. Recall that $\bar{\lambda}_2 = -\bar{\lambda}_1$. In this case, the function to be minimized is

$$\begin{aligned} & \ln Z(\lambda_1, \bar{\lambda}, \lambda_2, -\bar{\lambda}) + \langle \lambda_1 + \lambda_2, y_2 \rangle \\ &= \sum_{j=1}^{K_1} \ln \frac{1}{2} (e^{-g_{1,j}a} + e^{-g_{1,j}b}) + \sum_{j=1}^{K_2} \ln \frac{1}{2} (e^{-g_{2,j}a} + e^{-g_{2,j}b}) \\ &+ \sigma_{1,n}^2 (\lambda_1^2 + \bar{\lambda}^2) + \sigma_{2,n}^2 (\lambda_2^2 + \bar{\lambda}^2) + \langle \lambda_1 + \lambda_2, y_2 \rangle \end{aligned}$$

which is minimized over the region described in (12). Again, the optimal solutions (minimizing $\lambda_1^*, \bar{\lambda}^*, \lambda_2^*$) is to be found numerically. The estimated post-data is

$$dP_i^*(\xi) = \left(\prod_{j=1}^K (p_{i,j} \delta_a(d\zeta_j) + q_{i,j} \delta_b(d\zeta_j)) \right) \frac{e^{-\|v + \sigma_{i,n}^2 \mu_i^*\|^2 / 2 \sigma_{i,n}^2}}{(2\pi \sigma_{i,n}^2)^{N_i/2}} dv$$

for $i = 1, 2$ and where

$$p_{i,j} = \frac{e^{-ag_{i,j}}}{e^{-ag_{i,j}} + e^{-bg_{i,j}}} = 1 - q_{i,j}$$

from which the estimated parameters and residuals are given by

$$\beta_{i,j}^* = \frac{ae^{-ag_{i,j}} + be^{-bg_{i,j}}}{e^{-ag_{i,j}} + e^{-bg_{i,j}}} \quad \text{and} \quad \varepsilon_i^* = -\sigma_{i,n}^2 \mu_i^*$$

5. Empirical Example

We illustrate the applicability of our approach using an empirical application consisting of a small data set. The objective here is to demonstrate that our IT estimator is easy to apply and can be used for many different priors. The small sample performance of the IT-GME version of that estimator (uniform discrete priors) and detailed comparisons with other competing estimators is already shown in GMP and it falls outside the objectives of this note. The empirical example is based on one of the examples analyzed in GMP with data drawn from the March 1996 Current Population Survey. We estimated the wage-participation model for the subset of respondents in the labor market. Workers who are self-employed are excluded from the sample. Since the normal maximum likelihood estimator did not converge for that data [15], only results for the OLS, Heckman two-step, a semi-parametric estimator with a nonparametric selection mechanism due to [5], AP, and the different IT models developed here are reported [16]. To make our results comparable across the IT estimators, we use the empirical

standard deviations in all three cases and use supports between -100 and 100 for the IT-GME (uniform discrete priors) and the IT-Bernoulli case. In both the IT-Normal and IT-Bernoulli the priors used for the noise components are normal (as is shown in Section 4). Under these very similar specifications, we would expect all three IT examples to yield comparable estimates. Naturally, there are many other priors to choose from, but the objective here is just to show the flexibility and applicability of our approach.

We analyze a sample of 151 Native American females, of whom 65 are in the labor force. The wage equation covariates include years of education, a dummy for currently enrolled in school, potential experience ($\text{age} - \text{education} - 6$) and potential experience squared, a dummy for rural location, and a dummy for central city location. The covariates in the selection equation include all the variables in the wage equation and the amount of welfare payments received in the previous year, a dummy equal one for married, and the number of children. We use the three exclusion restrictions to identify the wage equation in the parametric and nonparametric two-step approaches.

Table 1. Estimates of the Native American wage equation (151 individuals; 65 in labor force).

	OLS	2-Step	AP	IT-GME	IT-Normal	IT-Bernoulli
Constant	1.073	1.771	NA	1.038	1.049	1.068
Education	0.055	0.043	0.044	0.054	0.056	0.055
Experience	0.038	0.023	0.038	0.038	0.038	0.038
Experience Squared	−0.001	−0.0005	−0.001	−0.001	−0.001	−0.001
Rural	0.214	0.268	0.332	0.210	0.215	0.214
Central City	−0.170	−0.091	−0.171	−0.186	−0.166	−0.169
Enrolled in School	−0.290	−0.471	−0.190	−0.301	−0.283	−0.288
λ		−0.461				
R^2	0.355	0.376	NA	0.343	0.355	0.354
MSPE	0.157	0.135	NA	0.147	0.144	0.144

Notes: Bold numbers reflect significantly different than zero at the 10% level

Table 1 presents the estimated coefficients for the wage equation. The R^2 and Mean Squared Prediction Error (MSPE) for each model are presented as well. All IT estimators outperform the other estimators in terms of predicting selection [17]. The estimated return to education is about 5% across all estimation methods, but only statistically significantly different from 0 for the OLS and the IT estimators. Though, all estimators have estimated parameters of the same magnitude and sign, only the OLS and the three reported IT estimates are statistically significantly different from zero in most cases.

6. Conclusion

In this short paper we develop a simple to apply, information-theoretic, method for analyzing nonlinear data with sample selection problem. Rather than using a likelihood approach or a semi-parametric approach we generalized further the IT-GME model of Golan, Moretti and Perloff (2004). Our model (i) allows for bounded and unbounded supports on all the unknown parameters, (ii) allows us to use a whole class of priors (continuous or discrete), (iii) is specified as a nonlinear concentrated entropy model, and (iv) is easy to apply. Like GMP our model works well even with

small data. This is shown in our empirical example. The extensions developed here mark a significant improvement on the GMP model and other IT, generalized entropy models.

A detailed set of sampling experiments comparing our IT method with all other competitors, under different data processes, will be done in future work.

Acknowledgement

We thank Enrico Moretti and Jeff Perloff for their comments on earlier versions of this work.

References and Notes

1. Heckman, J. Sample selection bias as a specification error. *Econometrica* **1979**, *47*, 153-161.
2. Manski, C.F. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *J. Econom.* **1985**, *27*, 313-333.
3. Cosslett, S.R. Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* **1981**, *51*, 765-782.
4. Han, A.K. Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *J. Econom.* **1987**, *35*, 303-316.
5. Ahn, H.; Powell, J.L. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *J. Econom.* **1993**, *58*, 3-29.
6. Golan, A.; Moretti E.; Perloff, J.M. A small sample estimation of the sample selection model. *Econom. Rev.* **2004**, *23*, 71-91.
7. Golan, A.; Judge, G.; Miller, D. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*; John Wiley & Sons: New York, NY, USA, 1996.
8. Golan, A.; Judge, G.; Perloff, J.M. Recovering information from censored and ordered multinomial response data. *J. Econom.* **1997**, *79*, 23-51.
9. Golan A. An information theoretic approach for estimating nonlinear dynamic models. *Nonlinear Dynamics Econom.* **2003**, *7*, 2.
10. Maddala, G.S. *Limited-Dependent and Qualitative Variables in Econometrics*; Cambridge University Press: Cambridge, UK, 1983.
11. Kullback, S. *Information Theory and Statistics*; John Wiley & Sons: New York, NY, USA, 1959.
12. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Statist.* **1951**, *22*, 79-86.
13. Gokhale, D.V.; Kullback, S. *The Information in Contingency Tables*; Marcel Dekker: New York, NY, USA, 1978.
14. Golan, A.; Gzyl, H. An information theoretic estimator for the linear model. *Working paper*, **2008**.
15. Due to the small size of the data and the large proportion of censored observations, none of the maximum likelihood methods would converge with all standard software.
16. For discussion of the data, detailed analyses and discussion of the different estimators as well as a detailed discussion of the AP [5] application, see GMP [6]. The 2-step and AP estimates are taken from that paper, Table 8.
17. To keep the Table simple and since these specific results are not of interested here, they are not presented.

Appendix

Proof of Theorem 2.1.

Proof: Recall (11)

$$\begin{aligned} & \text{Sup}\{\text{Sup}\{S(P_1, Q_1) + S(P_2, Q_2) \mid y_2 = B_2 E_{P_2}[\xi_2], \eta_1 = B_1 E_{P_1}[\xi_1]; \\ & \quad \bar{\eta}_2 = \bar{B}_2 E_{P_2}[\xi_2]; \bar{\eta}_1 = \bar{B}_1 E_{P_1}[\xi_1]\} \mid y_2 > \eta_1, \bar{\eta}_2 \leq \bar{\eta}_1\} \end{aligned} \quad (\text{A.1})$$

First, note that the inner problem in (A.1) is equivalent to

$$\ell(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) = \ln Z(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) + \langle \lambda_1, \eta_1 \rangle + \langle \bar{\lambda}_1, \bar{\eta}_1 \rangle + \langle \lambda_2, y_2 \rangle + \langle \bar{\lambda}_2, \bar{\eta}_2 \rangle$$

where λ_i and $\bar{\lambda}_i$ ($i=1, 2$) are the Lagrange multipliers associated with the data (9) and Z is the normalization factor of $dP_1 dP_2$:

$$\begin{aligned} Z(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) &= \iint_{C_1 \times C_2} e^{-\langle \lambda_1, B_1 \xi_1 \rangle} e^{-\langle \bar{\lambda}_1, \bar{B}_1 \xi_1 \rangle} e^{-\langle \lambda_2, B_2 \xi_2 \rangle} e^{-\langle \bar{\lambda}_2, \bar{B}_2 \xi_2 \rangle} dQ_1(\xi_1) dQ_2(\xi_2) \\ &= \int_{C_1} e^{-\langle \lambda_1, B_1 \xi_1 \rangle} e^{-\langle \bar{\lambda}_1, \bar{B}_1 \xi_1 \rangle} dQ_1(\xi_1) \int_{C_2} e^{-\langle \lambda_2, B_2 \xi_2 \rangle} e^{-\langle \bar{\lambda}_2, \bar{B}_2 \xi_2 \rangle} dQ_2(\xi_2) = Z_1(\lambda_1, \bar{\lambda}_1) Z_2(\lambda_2, \bar{\lambda}_2) \end{aligned}$$

Note that the inner *sup* is over (P_1, P_2) , and the outer *sup* is over the η 's in the region indicated within the $\{\cdot\}$. The basic idea here is to replace the inequalities appearing in problem (10) with equalities. Next, the dual-unconstrained model of this inner primal problem is the solution to

$$\inf_{\lambda} \left\{ \ell(\lambda_i, \bar{\lambda}_i) \mid \lambda_i \in \mathbb{R}^{|J|}, \text{ and } \bar{\lambda}_i \in \mathbb{R}^{N-|J|} \text{ for } i = 1, 2 \right\}$$

where $|J|$ is the number of observations where $y_{2i} > y_{1i}$. With this step, the equivalent dual model of the primal problem (A.1) is

$$\begin{aligned} & \text{Sup}\{\text{Inf}\{\ln Z(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) + \langle \lambda_1, \eta_1 \rangle + \langle \bar{\lambda}_1, \bar{\eta}_1 \rangle + \langle \lambda_2, y_2 \rangle + \langle \bar{\lambda}_2, \bar{\eta}_2 \rangle \mid \\ & \quad \lambda_i \in \mathbb{R}^{|J|}, \text{ and } \bar{\lambda}_i \in \mathbb{R}^{N-|J|}\} \mid y_2 > \eta_1, \bar{\eta}_2 \leq \bar{\eta}_1 \text{ for } i = 1, 2\} \end{aligned} \quad (\text{A.2})$$

Next, we rewrite the constraints for the outer problem. The constraint $y_2 > \eta_1$ is rewritten as $\eta_1 = y_2 - \zeta$, $\zeta > 0$, and the constraint $\bar{\eta}_2 \leq \bar{\eta}_1$ is written as $\bar{\eta}_2 \in \mathbb{R}^{N-|J|}$, $\bar{\eta}_1 = \bar{\eta}_2 + \bar{\zeta}$, $\bar{\zeta} \geq 0$. Model (A.2) becomes

$$\begin{aligned} & \text{Sup}\{\text{Inf}\{\ln Z(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) + \langle \lambda_1 + \lambda_2, y_2 \rangle - \langle \lambda_1, \zeta \rangle + \langle \bar{\lambda}_1 + \bar{\lambda}_2, \bar{\eta}_1 \rangle + \langle \bar{\lambda}_2, \bar{\zeta} \rangle \mid \\ & \quad \lambda_i \in \mathbb{R}^{|J|}, \text{ and } \bar{\lambda}_i \in \mathbb{R}^{N-|J|}\} \mid \zeta > 0, \bar{\eta}_1, \bar{\zeta} \geq 0\} \end{aligned}$$

Next, exchanging the *sup* and the *inf* operations we get

$$\begin{aligned} & \text{Inf}\{\ln Z(\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2) + \langle \lambda_1 + \lambda_2, y_2 \rangle + \text{Sup}\{-\langle \lambda_1, \zeta \rangle + \langle \bar{\lambda}_1 + \bar{\lambda}_2, \bar{\eta}_1 \rangle + \langle \bar{\lambda}_2, \bar{\zeta} \rangle \mid \\ & \quad \zeta > 0, \bar{\eta}_1, \bar{\zeta} \geq 0\} \mid \lambda_i \in \mathbb{R}^{|J|}, \text{ and } \bar{\lambda}_i \in \mathbb{R}^{N-|J|}\} \end{aligned}$$

To compute the inner supremum, note that

$$\begin{aligned}\text{Sup}\{\langle \lambda, \zeta \rangle \mid \zeta > 0\} &= \begin{cases} 0 & \text{if } -\lambda \in \mathbb{R}_+^d \\ \infty & \text{otherwise} \end{cases} \\ \text{Sup}\{\langle \lambda, \zeta \rangle \mid \bar{\zeta} \geq 0\} &= \begin{cases} 0 & \text{if } -\lambda \in \mathbb{R}_+^d \\ \infty & \text{otherwise} \end{cases}\end{aligned}$$

where $\mathbb{R}_+^d$ denotes the non-negative orthant in $\mathbb{R}^d$, the right hand side of the last identity is usually written as $I_{\mathbb{R}_+^d}(-\lambda)$ and $I_A(x)$ is defined as $I_A(x) = \begin{cases} 0 & \text{if } x \in A \\ \infty & \text{if } x \notin A \end{cases}$. The difference between the first and second problem is that in the first the supremum is reached only when $\lambda = 0$, whereas in the second it is reached at the boundary of $\mathbb{R}_+^d$. Similarly,

$$\text{Max}\{\langle \lambda, \zeta \rangle \mid \zeta \in \mathbb{R}^d\} = I_{\{0\}}(\lambda)$$

Noting that $\bar{\lambda}_1 = -\bar{\lambda}_2 \equiv \bar{\lambda}$, our MinMax problem (A.2) reduces to finding

$$\text{Inf}\left\{\ln Z(\lambda_1, -\bar{\lambda}, \lambda_2, \bar{\lambda}) + \langle \lambda_1 + \lambda_2, y_2 \rangle + I_{\mathbb{R}_+^d}(\lambda_1) + I_{\mathbb{R}_+^d}(\bar{\lambda}_1) \mid \lambda_1 \in \mathbb{R}^{|J|}, \lambda_2 \in \mathbb{R}^{|J|} \text{ and } \bar{\lambda} \in \mathbb{R}^{N-|J|}\right\}$$

which simplified to

$$\text{Inf}\left\{\ln Z(\lambda_1, -\bar{\lambda}, \lambda_2, \bar{\lambda}) + \langle \lambda_1 + \lambda_2, y_2 \rangle \mid \lambda_1 \in \mathbb{R}_+^{|J|}, \lambda_2 \in \mathbb{R}^{|J|} \text{ and } \bar{\lambda}_1 \in \mathbb{R}_+^{N-|J|}\right\}$$

which is (12). □