UC Berkeley
Working Paper Series

Title
GMM Estimation of a Maximum Distribution With Interval Data

Permalink
https://escholarship.org/uc/item/7jf5w1ht

Authors
Wu, Ximing
Perloff, Jeffrey M.

Publication Date
2005-03-01

GMM Estimation of a Maximum Entropy Distribution with Interval Data

Ximing Wu[*] and Jeffrey M. Perloff[**]

March 2005

Abstract

We develop a GMM estimator for the distribution of a variable where summary statistics are available only for intervals of the random variable. Without individual data, one cannot calculate the weighting matrix for the GMM estimator. Instead, we propose a simulated weighting matrix based on a first-step consistent estimate. When the functional form of the underlying distribution is unknown, we estimate it using a simple yet flexible maximum entropy density. Our Monte Carlo simulations show that the proposed maximum entropy density is able to approximate various distributions extremely well. The two-step GMM estimator with a simulated weighting matrix improves the efficiency of the one-step GMM considerably. We use this method to estimate the U.S. income distribution and compare these results with those based on the underlying raw income data.

[*] Department of Economics, University of Guelph; email: xiwu@uougelph.ca.

[**] Department of Agricultural and Resource Economics, University of California; email: perloff@are.berkeley.edu.

Introduction

Economists often want to estimate the distribution of a variable for which they have only summary statistics. Although we can use existing methods to recover the distribution when summary statistics for the entire distribution are available, often summary statistics are available only for intervals of the random variable. We introduce new methods to estimate a distribution when only limited information about intervals or ranges is available.

An important application of our technique is the estimation of income distributions as many government agencies report summary statistics for only some ranges of the income distribution: the income distribution is divided into a fixed number of intervals and only the share and sometimes the conditional mean of each interval are reported (see for example, Wu and Perloff, 2004 on the Chinese income distribution). Similarly, many governments provide only aggregated data on the distribution of firm size. For example, the U.S. government reports the market shares for only the 4, 8, 20, and 50 largest firms and the Herfindahl-Hirschman Index (sum of squared market shares) for the 50 largest firms.[1] Also, contingent valuation studies typically estimate the distribution of willingness-to-pay or willingness-to-accept based on answers to hypothetical dichotomous-choice questions. Consequently, estimating the underlying distribution is not possible using traditional approaches.

We propose a generalized method of moments (GMM) approach to estimate the underlying distribution based on summary statistics by intervals. Because we do not have individual data to calculate the weighting matrix for the GMM estimator, we simulate a weighting matrix based on consistent first-step estimates. We illustrate the properties of our

---

[1] Golan, Judge, and Perloff (1996) estimated firm size distribution based on these market concentration indices using maximum entropy techniques.

1

approach using Monte Carlo experiments and an experiment based on real-world data. Our experiments show that this GMM estimator with a simulated weighting matrix is substantially more efficient than the one-step GMM estimator and as efficient as the maximum likelihood estimator when it is feasible.

Because the functional form of the underlying distribution is usually unknown, we use a maximum entropy (maxent) density to approximate it. The maxent approach is a method to assign values to probability distributions based on limited information. Because the maxent density belongs to the exponential family, one can use some generalized moment conditions to summarize the distribution completely. These characterizing moments are sufficient statistics of the distribution. We propose a simple, yet flexible maxent density with a small number of characterizing moments. Our Monte Carlo simulations show the proposed maxent density is flexible enough to approximate closely various distributions that are skewed, fat-tailed, or even multi-modal.

In next section, we develop the GMM density estimator with a simulated weighting matrix based on interval summary statistics. Next, we introduce the maxent density approach to approximate the unknown underlying distribution. We then use Monte Carlo simulations to show that our approach works well. Finally, we apply this method to raw income data from the 1999 U.S. family income distribution and show that we can approximate the underlying distribution closely using only a limited amount of summary statistics by intervals. We find that our estimates successfully capture the features of the empirical distribution. The last section presents a summary and conclusions.

Estimation of Known Distributions

We first develop a GMM estimator of distributions based on interval data when the distribution is known. Let **x** be an *i.i.d.* random sample of size N from a continuous distribution $f(x;\boldsymbol{\theta})$ defined over a support I. We suppose that I is divided into K mutually exclusive and exhaustive intervals I_k, $k = 1,2,...,K$. We denote n_k and μ_k as the frequency and conditional mean of the sample **x** in the kth interval. In this section, we assume that we know the functional form of $f(x;\boldsymbol{\theta})$. We want to estimate $\boldsymbol{\theta}$, the vector of shape parameters.

One-Step GMM Estimator

If the only known information about $f(x; \boldsymbol{\theta})$ is n_k, one can estimate $f(x; \boldsymbol{\theta})$ using the maximum likelihood estimator (MLE). It is well known that the frequency table is distributed according to a multinomial distribution, and the likelihood function takes the form

$$L(\boldsymbol{\theta}) = N! \prod_{k=1}^{K} \frac{\left[P_k(\boldsymbol{\theta}) \right]^{n_k}}{n_k!},$$

where $N = \sum_{k=1}^{K} n_k$ and $P_k(\boldsymbol{\theta}) = \int_{I_k} f(x;\boldsymbol{\theta}) dx$.

Alternatively, one can use a GMM estimator. To ensure identification, the dimension of $\boldsymbol{\theta}$ must be no greater than K. Denote $\mathbf{m} = [m_1, m_2, ..., m_K]'$ where

$$m_k = P_k(\boldsymbol{\theta}) - s_k, \quad k = 1, 2, ..., K$$

and $s_k = n_k / N$. The objective function of the one-step GMM, which we call GMM$_1$, is

$$Q_{GMM_1}(\boldsymbol{\theta}) = \mathbf{m}(\boldsymbol{\theta})' \mathbf{m}(\boldsymbol{\theta}).$$

If the conditional mean for each group is also known, it is impractical to use the MLE as the likelihood function is too complicated. However, we can still use the GMM estimator. The moment conditions now take the form

$$m_k = \begin{bmatrix} P_k(\boldsymbol{\theta}) - s_k \\ \mu_k(\boldsymbol{\theta}) - \mu_k \end{bmatrix}, \quad k = 1, 2, ..., K,$$ (1)

where $\mu_k(\boldsymbol{\theta}) = \int_{I_k} xf(x; \boldsymbol{\theta}) dx$.

We can easily apply this method of moments approach to accommodate other forms of information by intervals, such as higher-order moments, order statistics, Gini indexes, and so on. Therefore, the GMM_1 approach to estimating a distribution based on interval data is much more flexible than the MLE.

Two-Step GMM Estimator with Simulated Weighting Matrix

Although the GMM_1 can incorporate more information than the MLE, it is generally not efficient unless its optimal weighting matrix $\boldsymbol{\Omega}$ coincidentally equals the identity matrix. To gain efficiency, one can estimate a two-step GMM estimator (GMM_2), which is obtained by minimizing

$$Q_{GMM_2}(\boldsymbol{\theta}) = \mathbf{m}(\boldsymbol{\theta})'\boldsymbol{\Omega}\mathbf{m}(\boldsymbol{\theta}).$$

Denote the variance-covariance matrix of the moment conditions $\mathbf{m}(\boldsymbol{\theta})$ as $\mathbf{W}$, the optimal choice of the weighting matrix is $\boldsymbol{\Omega} = \mathbf{W}^{-1}$.

However because we have only interval data, we are not able to calculate the variance-covariance matrix of the moment conditions and hence cannot use the GMM_2 estimator. Instead, we propose a method of simulating the weighting matrix based on the first-step estimates. Since the one-step GMM_1 estimate is consistent, so is the simulated weighting matrix based on

4

$f\left(x;\hat{\mathbf{\theta}}_{GMM_1}\right)$. Therefore, like the conventional two-step GMM$_2$, our alternative simulation approach, GMM$_s$ is asymptotically efficient. Our simulations suggest that the GMM$_s$ estimator obtains the same level of efficiency as do the MLE and the more data-intensive GMM$_2$ procedures.

We first consider the case where we lack the conditional means, so that we can use the MLE approach practically, and then we consider the more complicated case where we also have additional information such as conditional means. Given the density function $f\left(x;\mathbf{\theta}\right)$, the MLE based on the multinomial distribution is efficient when only interval share, or the frequency table, is known. For the GMM$_2$ estimator, we need to calculate the weighting matrix. Because a frequency table is the sufficient statistics of a multinomial distribution, we do not need x to compute its variance covariance matrix, which is given by[2]

$$\mathbf{W} = \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & s_K \end{bmatrix}. \tag{2}$$

The resulting GMM estimator is a minimum chi-square estimator.

Instead of using the asymptotically optimal weighting matrix, one can use a simulated weighting matrix. To obtain the simulated weighting matrix, we do the following:

1. Draw an *i.i.d.* random sample $\mathbf{x}^*$ of size N from the GMM$_1$ estimate $f\left(x;\hat{\mathbf{\theta}}_{GMM_1}\right)$;

2. Group $\mathbf{x}^*$ in the same way as the original interval data and calculate the frequency table $\mathbf{s}^*$;

---

[2] The derivation of this variance covariance matrix can be found in the typical treatment of chi-square goodness-of-fit test of grouped data (see, for example, Agresti, 1990).

3. Calculate the variance-covariance matrix of $\mathbf{s}^*$ as in Equation (2).

This procedure is repeated B times and the weighting matrix is the average of the simulated weighting matrices:

$$\mathbf{\Omega}^* = \frac{1}{B}\sum_{b=1}^{B}\left(\mathbf{W}_{(b)}^*\right)^{-1}.$$ (3)

Our new GMM estimator with simulated weighting matrix (GMM$_s$) is then defined by minimizing

$$Q_{GMM_s} = \mathbf{m}(\mathbf{\theta})'\,\mathbf{\Omega}^*\mathbf{m}(\mathbf{\theta}).$$

We now turn to the more complicated case where the conditional mean of each interval is also used. Now, we are not able to calculate the asymptotic weighting matrix for the GMM$_2$, which requires knowledge of the full sample. Nonetheless, the GMM$_s$ is still applicable and should be more efficient than the GMM$_1$. In this case, we obtain the simulated weighting matrix by using an alternative three-step procedure:

1. Draw an *i.i.d.* random sample $\mathbf{x}^*$ of size N from the GMM$_1$ estimate $f\left(x;\hat{\mathbf{\theta}}_{GMM_1}\right)$;

2. Group $\mathbf{x}^*$ in the same way as the original grouped data and calculate the frequency table $\mathbf{s}^*$ and conditional mean $\mathbf{\mu}^*$;

3. Define $D_k(x)=1$, $x\in I_k$ and 0 otherwise, and

$$\mathbf{w}_k^* = \begin{bmatrix} \sum_{t=1}^{N}D_k\left(x_t^*\right) & \sum_{t=1}^{N}D_k\left(x_t^*\right)x_t^* \\ \sum_{t=1}^{N}D_k\left(x_t^*\right)x_t^* & \sum_{t=1}^{N}D_k\left(x_t^*\right)\left(x_t^*\right)^2 \end{bmatrix}.$$ (4)

then calculate

6

$$\mathbf{W}^* = \begin{bmatrix} \left(\mathbf{w}_1^*\right) & 0 & \cdots & 0 \\ 0 & \left(\mathbf{w}_2^*\right) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \left(\mathbf{w}_K^*\right) \end{bmatrix}.$$

This procedure is repeated B times and the simulated weighting matrix is then calculated as in Equation (3). Because the direct calculation of the weighting matrix in Equation (4) requires individual data, the conventional two-step GMM_2 is infeasible when only interval summary statistics are available.

Large Sample Properties

The proposed GMM estimators share the same large sample properties of the standard GMM estimator (see for example, Newey and McFacdden, 1994).

Assumptions: For $\hat{Q}_n(\boldsymbol{\theta}) = \hat{\mathbf{m}}_n(\boldsymbol{\theta})'\hat{\boldsymbol{\Omega}}\hat{\mathbf{m}}_n(\boldsymbol{\theta})$,

(a) $\boldsymbol{\theta}_0 \in \text{interior}(\boldsymbol{\Theta})$, which is compact, is the unique solution to the moment condition $E\left[\mathbf{m}(x_i, \boldsymbol{\theta})\right] = 0$;

(b) $\mathbf{m}_n(\boldsymbol{\theta})$ is continuous on $\boldsymbol{\Theta}$ and for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ it is twice continuously differentiable in a neighborhood $\mathbf{N}$ of $\boldsymbol{\theta}_0$;

(c) $E\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \| \mathbf{m}(x, \boldsymbol{\theta}) \|\right] < \infty$;

(d) $\hat{\boldsymbol{\Omega}} \xrightarrow{P} \boldsymbol{\Omega}$, $\boldsymbol{\Omega}$ is positive semi-definite;

(e) $\sqrt{n}\hat{\mathbf{m}}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N(0, \mathbf{W})$;

(f) $\mathbf{G}(\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_0$ and $\sup_{\boldsymbol{\theta} \in \mathbf{N}} \| \nabla_{\boldsymbol{\theta}}\hat{\mathbf{m}}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta}) \| \xrightarrow{P} 0$, where $\nabla_{\boldsymbol{\theta}}\hat{\mathbf{m}}_n(\boldsymbol{\theta})$ is the gradient of $\hat{\mathbf{m}}_n(\boldsymbol{\theta})$;

(g) $\mathbf{G}(\mathbf{\theta}_0)'\mathbf{\Omega G}(\mathbf{\theta}_0)$ is nonsingular;

(h) $\mathbf{\Omega} = \text{plim}(\hat{\mathbf{\Omega}}) = \mathbf{W}^{-1}$.

Under assumptions (a)-(g), all GMM estimators discussed in this study are asymptotically normal with

$$\sqrt{n}(\hat{\mathbf{\theta}} - \mathbf{\theta}_0) \xrightarrow{d} N\left[0, (\mathbf{G'\Omega G})^{-1}\mathbf{G'\Omega'W\Omega G}(\mathbf{G'\Omega G})^{-1}\right].$$

For the GMM$_1$, $\mathbf{\Omega}$ is the identity matrix and the asymptotic variance covariance matrix is $(\mathbf{G'G})^{-1}\mathbf{G'WG}(\mathbf{G'G})^{-1}$. Because it satisfies assumption (a)-(h), the GMM$_2$ (when it is feasible), is asymptotically efficient with

$$\sqrt{n}(\hat{\mathbf{\theta}} - \mathbf{\theta}_0) \xrightarrow{d} N\left[0, (\mathbf{G'W}^{-1}\mathbf{G})^{-1}\right].$$

By construction, the simulated weighting matrix $\mathbf{\Omega}^*$ of the GMM$_s$ based on first-step consistent estimate converges asymptotically to the optimal weighting matrix $\mathbf{W}^{-1}$. Therefore, the GMM$_s$ also satisfies assumption (a)-(h) and attains the same asymptotic efficiency as that of the GMM$_2$.

Approximation of Unknown Distributions via Maximum Entropy Densities

Very often the functional form of the underlying density is unknown, and various parametric densities are used to approximate the unknown density. For example, the log-normal density is a popular choice to model the income distribution. Instead of relying on a limited collection of known densities, we propose to use the maximum entropy (maxent) densities for density approximation. When the maxent density is identical to the unknown underlying distribution, our estimator is fully efficient. When it encompasses the underlying distribution, the estimator is nearly efficient as the efficiency loss of maxent density estimation due to a small number of redundant parameters is negligible (Wu and Stengos, 2004). If the maxent density is a

close approximation to the unknown distribution, then one can expect the efficiency of the GMM estimator to be close to that of the MLE.[3] Fortunately, our experiments show that the proposed specification of maxent density is flexible enough to approximate closely various distributions that are skewed, fat-tailed, or multi-modal.

Background and Generalization to Interval Data

We use a maxent approach to approximate unknown distributions. The principle of maximum entropy is a general method to assign values to probability distributions using limited information. This principle, introduced by Jaynes in 1957, states that one should choose the probability distribution, consistent with the given constraints, that maximizes Shannon's entropy. According to Jaynes (1957), the maximum entropy distribution is "uniquely determined as the one which is maximally noncommittal with regard to missing information, and that it agrees with what is known, but expresses maximum uncertainty with respect to all other matters."

Maximizing entropy subject to various side conditions is well known in the literature as a method of deriving the forms of minimal information prior distribution (see for example, Jaynes 1968 and Zellner 1977). The maxent density $f(x)$ is obtained by maximizing Shannon's information entropy

$$W = -\int_I f(x) \log f(x)\, dx$$

subject to K known moment conditions for the entire range of the distribution

---

[3] Using the Kullback-Leibler distance to quantify the closeness or discrepancy between the maxent density and the unknown distribution, one can show that the GMM estimator converges to the MLE asymptotically as the Kullback-Leibler distance approaches zero (see for example Golan, Judge, and Miller, 1996, for a discussion of the duality between these two approaches for likelihood function defined over the exponential family distribution).

9

$$\int_I f(x)\,dx = 1,$$

$$\int_I g_i(x) f(x)\,dx = v_i,$$

where $i = 1, 2,\ldots, M$ indexes the characterizing moments, v_i, and their functional forms, $g_i(x)$.

Here $g_i(x)$ is continuous and at least twice differentiable.

We can solve this optimization problem using Lagrange's method, which leads to a unique global maximum entropy. The solution takes the form

$$f(x;\boldsymbol{\theta}) = \exp\left(-\theta_0 - \sum_{i=1}^{M} \theta_i g_i(x)\right),$$

where θ_i is the Lagrange multiplier for the ith moment constraint and

$\theta_0 = \log[\int \exp\left(-\sum_{i=1}^{M} \theta_i g_i(x)\right)]$ is the normalization term that ensures the density integrates to

one.

Zellner and Highfield (1988) and Wu (2003) discuss the estimation of maxent density subject to moment constraints for the entire distribution. Generally the maxent density estimation method has no analytical solution. To solve for the Lagrange multipliers, we use Newton's method to iteratively update

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + \mathbf{G}^{-1}\mathbf{m},$$

where $\mathbf{G}$ is the $(M+1)$ by $(M+1)$ Hessian matrix of the form

$$G_{ij} = \int_I g_i(x) g_j(x) f(x;\boldsymbol{\theta})\,dx, \quad 0 \le i, j \le M,$$

and

$$m_i = G_{i0} - v_i, \quad 0 \le i \le M.$$

This maximum entropy method is equivalent to a maximum likelihood approach where the likelihood function is defined over the exponential distribution and therefore consistent and efficient.

In this study, we extend the classical maxent density method to deal with interval data. Given the moment conditions on interval share and conditional mean as in Equation (1), the GMM_1 estimator for a maxent density $f(x; \theta)$ can be solved by iteratively updating

$$\theta^{(1)} = \theta^{(0)} + \left(\mathbf{G}'\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{m},$$

where

$$m_k = \begin{bmatrix} \int_{I_k} f\left(x;\theta^{(0)}\right)dx - s_k \\ \int_{I_k} xf\left(x;\theta^{(0)}\right)dx - \mu_k \end{bmatrix}$$

and the $2K$ by M matrix $\mathbf{G}$ consists of the K stacked submatrices $\mathbf{G}^{(k)} = \left[\mathbf{G}_1^{(k)}, \mathbf{G}_2^{(k)}, ..., \mathbf{G}_M^{(k)}\right]$, where

$$\mathbf{G}_i^{(k)} = \begin{bmatrix} \int_{I_k} g_i(x) f\left(x;\theta^{(0)}\right)dx \\ \int_{I_k} xg_i(x) f\left(x;\theta^{(0)}\right)dx \end{bmatrix}, \quad i = 1, 2, ..., M.$$

To improve efficiency, we can use the GMM_s, whose weighting matrix $\mathbf{\Omega}^*$ is simulated from the consistent one-step GMM_1. The parameters are then obtained using the updating formula

$$\theta^{(1)} = \theta^{(0)} + \left(\mathbf{G}'\mathbf{\Omega}^*\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{\Omega}^*\mathbf{m}.$$

Flexible Specification of Maximum Entropy Density

Barron and Sheu (1991) characterized the maxent density alternatively as an approximation of the log density by some basis functions, such as polynomials, trigonometric

series, or splines. They showed that the estimator does not depend on the choice of basis

function. One can measure how close an estimated distribution $\hat{f}(x)$ is to the true distribution

$f(x)$ using the Kullback-Leibler distance measure,

$$D = \int_I f(x) \log \frac{f(x)}{\hat{f}(x)} dx .$$

Under some regularity conditions, the maxent density estimate converges to the underlying

density, in terms of the Kullback-Leibler distance, as the number of moment conditions increases

with sample size.[4]

Theoretically, one can approximate an unknown distribution arbitrarily well using the

maxent density. In practice, often only a small number of moment conditions are used because

the Hessian matrix quickly approaches singularity as the number of moment conditions increases.

Nonetheless, one can closely approximate distributions of various shapes using the maxent

densities subject to a few moment conditions. In the following simulations, we use a simple, yet

flexible maxent density:

$$f(x;\boldsymbol{\theta}) = \exp\left(-\theta_0 - \theta_1 x - \theta_2 x^2 - \theta_3 \arctan(x) - \theta_4 \log(1+x^2)\right), \qquad (5)$$

where arctan(x) is the inverse tangent function.[5] Density function (5) nests the normal as a

special case when $\theta_3 = \theta_4 = 0$. We use arctan(x) and log(1+x^2) to capture deviations from the

---

[4] Tagliani (2003) showed that $V \le 3\left[-1+(1+4D/9)^{1/2}\right]^{1/2}$, where $V = \int |f(x) - \hat{f}(x)| dx$ is the commonly-used variation measure. Hence, convergence in the Kullback-Leibler distances implies convergence in the variation measure.

[5] The arctan(x) is also used by Bera and Park (2004) in the maximum entropy estimation of the ARCH model.

normal distribution. Both terms are of lower order than x^2 and are therefore not as sensitive to outliers.[6]

The arctan(x) is an odd function and is able to capture skewness and other deviations from the bell shape of symmetric distribution, such as that of normal or t distribution. Because its range is restricted between $-\pi/2$ and $\pi/2$, it limits the influence of potential outliers. Therefore, the arctan(x) is more resistant to outliers compared to the unbounded x^3 for measuring skewness.

The term $\log(1+x^2)$ is introduced to accommodate fat tails. Note that the fat-tailed student t distribution may be described as a maxent density with a characterizing moment $\log(1+x^2/v)$, where v is the degrees of freedom. Usually the degrees of freedom parameter v is unknown and direct estimation of v places this unknown parameter on both sides of the moment constraint,

$$\int_I \log\left(1+x^2/v\right)\exp\left(-\theta_0 - \theta_1 \log\left(1+x^2/v\right)\right)dx = \frac{1}{n}\sum_{t=1}^{n}\log\left(1+x_t^2/v\right),$$

in the maxent optimization problem, which results in a difficult saddle point problem. Instead, we choose to use a linear combination of x^2 and log $(1+x^2)$ to approximate log $(1+x^2/v)$. When there is one degree of freedom, or the distribution is Cauchy, log$(1+x^2)$ characterizes the density; on the other extreme, when the degrees of freedom goes to infinity, the t distribution approximates the normal distribution and x and x^2 characterize the density.

---

[6] By using higher order polynomials in the exponent, we can obtain alternative maxent densities. However, higher sample moments are more sensitive to outliers and consequently, so are the density estimators using these higher moments. Also, the sample moment ratios, such as skewness and kurtosis, are restricted by the sample size (Dalén, 1987).

To examine how well x^2 and $\log(1+x^2)$ approximate $\log(1+x^2/v)$, we use the ordinary least squares to regress $\log(1+x^2/v)$ on x^2, $\log(1+x^2)$ and a constant term. Because all functions involved are even, we only look at x on the positive real line. In the experiment, we set x as the vector of all the integers within [1, 10,000]. For an arbitrary integer v within [1, 100], the R^2 is always larger than 0.999, indicating that $\log(1+x^2/v)$ can be well approximated by x^2 and $\log(1+x^2)$.

Monte Carlo Simulations

We use Monte Carlo simulations to investigate the numerical properties of the proposed estimators. In each of our experiments, we set the sample size at 250 or 500 and repeat the experiment 1,000 times. The randomly generated sample is divided into six intervals, separating at the 10th, 25th, 50th, 75th and 90th percentile of the underlying population.[7] The subsequent share of each interval ranges from 10% to 25%. If all the intervals have equal shares, the optimal weighting matrix, Equation (2), will be proportional to the identity matrix, rendering all three estimators asymptotically equivalent. Therefore, to investigate the efficiency gain of the two-step GMM_2 and GMM_s over the GMM_1, we generate intervals with different shares. For the GMM_s, we draw 300 random samples from the first-step GMM_1 estimate to calculate the simulated weighting matrix.

In the first experiment, we generate the sample from the standard normal distribution and assume the functional form is known. When only a frequency table is used in the estimation, we can directly compare the two-known, asymptotically efficient estimators, MLE and GMM_2, to

---

[7] Very often, the underlying sample size of grouped summary statistics, such as those for household income survey, is well above 500. The number of groups is also usually larger than six.

GMM_s. For each experiment, we also test the hypothesis that the random sample of individual data is generated according to the estimated density using the two-sided Kolmogorov-Smirnov (KS) test. The top panel of Table 1 reports the mean and standard errors of the integrated mean squared errors (MSE) between the estimated density and the underlying true density, and the percentage of the Kolmogorov-Smirnov test that is rejected at the 5% significance level. [8]

As expected, the GMM_2, which uses the optimal asymptotic weighting matrix, is more efficient than the GMM_1. For a sample size of 250, the mean of the MSE is 1.293 for the GMM_2 and 1.402 for the GMM_1. The GMM_s, which replaces the asymptotic weighting matrix with a simulated weighting matrix, performs as well (its average MSE is 1.216) as the GMM_2, which has a directly calculated weighting matrix. Further, both the two-step GMM estimators are essentially as efficient as the MLE (its average MSE is 1.212, and GMM_s has a lower standard error of the MSE than the MLE has). In the four experiments, the null hypothesis that the random sample in question is generated according to the estimated maxent density is rejected at the 5% significance level at most once out of a thousand experiments. The simulation results with a sample size of 500 follow a similar general pattern, except that the MSE is considerably smaller (on average, the MSE is reduced by 48%).

When we also use the conditional mean of each interval in the estimation, only the GMM_1 and the GMM_s are feasible. The results are reported in the bottom panel of Table 1. As we would expect, incorporating extra information improves the efficiency of the estimates. For a sample size of 250, the average MSE drops from 1.216 to 1.007 for the GMM_s (the corresponding drop is 1.402 to 1.164 for the GMM_1). Consistent with the case when we only use frequency information, the two-step GMM_s is more efficient than the GMM_1 (the average MSE

---

[8] All the MSE numbers reported in Table (1)-(4) have been multiplied by 1,000.

is 1.007 for the GMM$_1$ and 1.164 for the GMM$_s$). The KS goodness-of-fit test is rejected no more than four times.

In the next three experiments, we relax the assumption that we know the functional form and use the maxent density to approximate the unknown underlying distributions. We consider three distributions: (i) the skewed-normal distribution with shape parameter one, which is mildly skewed;[9] (ii) the skewed-normal distribution with shape parameter three, which is very skewed; and (iii) a half-half mixture of two normals, N(-3,1) and N(3,1), which is bi-modal.

The simulation results are reported in Table 2, 3 and 4. The general pattern is very close to the first experiment where we know the true underlying distribution. The two-step GMM improves on the one-step GMM. On average, the efficiency gain of the GMM$_s$ relative to the GMM$_1$ is 7% when we only use frequency information and 40% when we also use conditional means with a sample size of 250. The corresponding improvement is 5% and 47% respectively for a sample size of 500. The two-step GMM is as efficient as the MLE where it is feasible. Also, the goodness-of-fit test does not reject the null hypothesis in most cases, especially when the more efficient two-step estimators are used.

We note that incorporating additional information in an inefficient way does not necessarily improve the estimates. The performance of the GMM$_1$ that use both the frequency table and conditional means is sometimes slightly worse than that of the GMM$_1$ based on only the frequency table. In contrast, when we use the two-step GMM with simulated weighting matrix, GMM$_s$, incorporating extra information into the estimation process always improves the estimates in our experiments. For example in the experiment with bi-modal normal mixture

---

[9] The skew-normal distribution is defined by the density function $2\phi(x)\Phi(\alpha x)$, where $\phi(x)$ and $\Phi(x)$ is the density and distribution function of the standard normal respectively and α is the shape parameter determining the degree of skewness.

distribution with a sample size of 250, the average MSE of GMM_1 is 2.175 when we use only the frequency table, but increases to 2.425 when we also incorporate conditional means. In contrast, for GMM_s, the average MSE decreases from 2.044 to 1.055 by 48%, reflecting the benefit of additional information.

Our experiments across sample sizes and underlying distributions show that (i) the two-step GMM estimators are as efficient as the MLE when it is feasible; (ii) the GMM_s estimator with a simulated weighting matrix is as efficient as its counterpart with a asymptotically optimal weighting matrix (both the average MSE and standard error of the MSE are smaller in all but one experiment); (iii) incorporating extra information in an efficient way improves the estimator; (iv) in more than 99% of experiments, the goodness-of-fit test does not reject the hypothesis that the sample in question is generated according to the estimated maxent density, indicating that the proposed maxent density specification is flexible enough to approximate density function of the various shapes considered in this study.

Empirical Application

In this section, we use the proposed method to estimate the U.S. income distribution using data from the 2000 U.S. Current Population Survey (CPS) March Supplement. We draw a random sample of 5,000 observations from the CPS and divide the sample into 6 intervals, separating at the 10th, 25th, 50th, 75th and 90th sample percentile. We then estimate the underlying distribution using: (i) only the frequency table; (ii) the frequency table and the conditional means of each interval. We only report the estimates from the GMM_1 and GMM_s, which are feasible for both cases.

Based on the Kolmogorov-Smirnov test statistics between the sample and estimated densities (first column of Table 5), we cannot reject at the 1% significance level the hypothesis

17

that the sample is distributed according to the estimated distribution for all four estimates. Moreover, two commonly used inequality measures, the Gini index (second column) and interquantile-range (third column), from the sample and the estimated densities, are extremely close. For example, the "true" Gini from the individual data is 0.4130 and the Gini measures based on the GMM_s are 0.4128 (frequency only) and 0.4136 (frequency and conditional mean). The corresponding interquartile-range statistics are 0.5110, 0.5007, and 0.5110.

We can also compare the estimated densities directly using graphs. Figure 1 plots the estimated densities from the GMM_s with and without the conditional mean information against the histogram of the full sample. Clearly, both estimates successfully capture the shape of the empirical distribution.

Conclusions

We develop a GMM estimator to estimate the distribution of a variable when the only data available are summary statistics by intervals. Because no data at the individual level are available to calculate the weighting matrix for the GMM estimator, we propose a simulated weighting matrix based on consistent first-step estimates. When the functional form of the underlying distribution is unknown, we use a simple yet flexible maximum entropy density to approximate it.

We use Monte Carlo simulation experiments to illustrate that our estimated densities based on interval data may approximate the underlying distribution with high precision. In our experiments, the two-step GMM estimator with simulated weighting matrix is as efficient as MLE (where it is feasible) and substantially more efficient than the one-step GMM estimator. Moreover, our proposed maximum entropy density is able to approximate various distributions that are skewed, fat-tailed, or multi-modal.

We employ the proposed method to estimate the 1999 U.S. income distribution from interval data and compare the results with the underlying raw income data from the Current Population Survey. Our estimates successfully capture the features of the empirical distribution.

References

Agresti, A., 1990, Categorical data analysis (John Wiley and Sons, New York).

Barron, A.R. and C. Sheu, 1991, Approximation of density functions by sequences of exponential families, Annals of Statistics 19, 1347-69.

Bera, A. and S.Y. Park, 2004, Maximum entropy autoregressive conditional heteroskedasticity model, working paper.

Dalén, J., 1987, Algebraic bounds on standardized sample moments, Statistics and Probability Letters 5, 329-31.

Golan, A., G. Judge and D. Miller, 1996, Maximum entropy econometrics: robust estimation with limited data (John Wiley and Sons, New York).

Golan, A., G. Judge and J.M. Perloff, 1996, Estimating the size distribution of firms using government summary statistics, Journal of Industrial Economics 44, 69-80.

Jaynes, E. T., 1957, Information theory and statistical mechanics, Physics Review 106, 620-30.

Jaynes, E.T., 1968, Prior possibilities, IEEE Transactions on Systems Science and Cybernetics SSC-4, 227-241.

Newey, W.K. and D. McFadden, 1994, Large sample estimation and hypothesis testing, in: R.F. Eagle and D. McFadden, eds., Handbook of Econometrics, Vol. 4 (North-Holland, Amsterdam).

Tagliani, A., 2003, A note on proximity of distributions in terms of coinciding moments, Applied Mathematics and Computation 145, 195-203.

Wu, X., 2003, Calculation of maximum entropy densities with application to income distribution, Journal of Econometrics 115, 347-54.

Wu, X. and J.M. Perloff, 2004, China's income distribution and inequality, working paper.

Wu, X. and T. Stengos, 2004, Partially adaptive estimation via maximum entropy densities, working paper.

Zellner, A., 1977, Maximal data information prior distribution, in: A. Aykac and C. Brumat, eds., New Methods in the Applications of Bayesian Methods (North-Holland, Amsterdam).

Zellner, A. and R.A. Highfield, 1988, Calculation of maximum entropy distribution and approximation of marginal posterior distributions, Journal of Econometrics 37, 195-209.

Table 1. Summary statistics of results for normal distribution

		MLE	GMM$_1$	GMM$_2$	GMM$_s$
Frequency only					
Average MSE	N=250	1.212	1.402	1.293	1.216
	N=500	0.623	0.712	0.662	0.639
S.E. of MSE	N=250	1.237	1.463	1.260	1.174
	N=500	0.628	0.713	0.658	0.663
KS test	N=250	0	0	0	0.001
	N=500	0	0	0	0
Frequency and Mean					
Average MSE	N=250		1.164		1.007
	N=500		0.573		0.504
S.E. of MSE	N=250		1.168		1.009
	N=500		0.566		0.499
KS test	N=250		0.003		0.002
	N=500		0.004		0

Table 2. Summary statistics of results for skewed normal distribution ($\alpha = 1$)

		MLE	GMM$_1$	GMM$_2$	GMM$_s$
Frequency only					
Average MSE	N=250	6.810	7.783	7.149	6.922
	N=500	3.201	3.547	3.256	3.201
S.E. of MSE	N=250	6.665	9.129	7.817	7.488
	N=500	3.060	3.403	2.934	2.854
KS test	N=250	0.045	0.059	0.047	0.043
	N=500	0.016	0.023	0.027	0.022
Frequency and Mean					
Average MSE	N=250		3.510		2.777
	N=500		2.585		1.522
S.E. of MSE	N=250		3.105		2.509
	N=500		2.344		1.290
KS test	N=250		0		0
	N=500		0		0

Table 3. Summary statistics of results for skewed normal distribution ($\alpha = 3$)

		MLE	GMM_1	GMM_2	GMM_s
Frequency only					
Average MSE	N=250	6.671	6.657	6.433	6.310
	N=500	3.360	3.623	3.545	3.516
S.E. of MSE	N=250	6.072	6.554	6.006	5.860
	N=500	3.166	3.239	3.178	3.148
KS test	N=250	0	0.002	0.002	0.002
	N=500	0.001	0.002	0.003	0.002
Frequency and Mean					
Average MSE	N=250		6.078		3.542
	N=500		3.179		1.768
S.E. of MSE	N=250		5.288		3.239
	N=500		2.779		1.541
KS test	N=250		0		0
	N=500		0		0

Table 4. Summary statistics of results for mixed normal distribution

		MLE	GMM_1	GMM_2	GMM_s
Frequency only					
Average MSE	N=250	2.082	2.175	2.099	2.044
	N=500	1.106	1.177	1.143	1.131
S.E. of MSE	N=250	1.904	1.916	1.796	1.753
	N=500	1.011	1.085	1.048	1.040
KS test	N=250	0	0	0	0
	N=500	0	0.001	0.002	0.002
Frequency and Mean					
Average MSE	N=250		2.425		1.055
	N=500		1.342		0.593
S.E. of MSE	N=250		2.299		0.799
	N=500		1.161		0.401
KS test	N=250		0.002		0
	N=500		0.004		0

Table 5. Estimation results for U.S. Income Distribution

	KS	Gini	I-Q Range
Sample		0.4130	0.5110
Frequency only			
GMM_1	0.0104	0.4106	0.5000
GMM_s	0.0092	0.4128	0.5007
Frequency and Means			
GMM_1	0.0104	0.4148	0.5011
GMM_s	0.0089	0.4136	0.5110

KS: Kolmogorov-Smirnov statistics
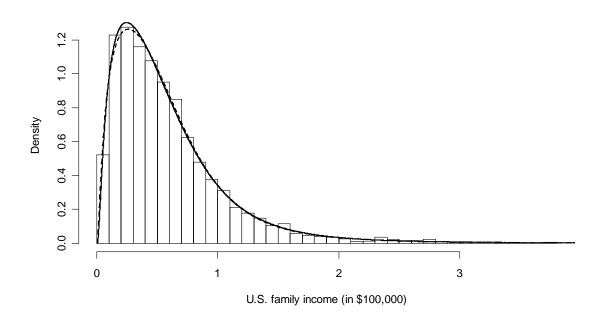I-Q Range: inter-quartile range (in $100,000).

Figure 1. Estimated densities (solid: GMM$_s$ based on frequency table; dotted: GMM$_s$ based on frequency table and conditional means) and a histogram based on individual data.