Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 115 (2003) 347–354
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

Calculation of maximum entropy densities with application to income distribution

Ximing Wu*

Department of Agricultural and Resource Economics, University of California at Berkeley, Berkeley, CA 94720, USA, and Department of Economics, University of Guelph, 'Guelph, Ont., Canada

Accepted 6 February 2003

Abstract

The maximum entropy approach is a flexible and powerful tool for density approximation. This paper proposes a sequential updating method to calculate the maximum entropy density subject to known moment constraints. Instead of imposing the moment constraints simultaneously, the sequential updating method incorporates the moment constraints into the calculation from lower to higher moments and updates the density estimates sequentially. The proposed method is employed to approximate the size distribution of U.S. family income. Empirical evidence demonstrates the efficiency of this method.

JEL classification: C4; C6; D3

Keywords: Maximum entropy; Density estimation; Sequential updating; Income distribution

0. Introduction

A maximum entropy (maxent) density can be obtained by maximizing Shannon's information entropy measure subject to known moment constraints. 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."

The maxent approach is a flexible and powerful tool for density approximation, which nests a whole family of generalized exponential distributions, including the exponential, Pareto, normal, lognormal, gamma, beta distribution as special cases.

* Tel.: +1-510-642-8179; fax: +1-510-643-8911.
E-mail address: ximing@are.berkeley.edu (X. Wu).

0304-4076/03/$ - see front matter
doi:10.1016/S0304-4076(03)00114-3

The maxent density has found some applications in econometrics. For example, see Zellner (1997) and Zellner and Tobias (2001) for the Bayesian method of moments, which uses the maxent technique to estimate the posterior density of parameters of interest; and Buchen and Kelly (1996), Stutzer (1996) and Hawkins (1997) for some applications in finance.

Despite its versatility and flexibility, the maxent density has not been widely used in empirical studies. One possible reason is that there is generally no analytical solution for the maxent density problem and the numerical estimation is rather involved. In this study, I propose a sequential updating method for the calculation of maxent densities. Compared to the existing studies that consider the estimation of the maxent density subject to just a few moment constraints, the proposed method is able to calculate the maxent density associated with a much higher number of moment constraints. This method is used to approximate the size distribution of U.S. family income distribution.

1. The maxent density

The maxent density is typically obtained by maximizing Shannon's entropy (defined relative to uniform measure),

$$W = \int -p(x)\log p(x)\,\mathrm{d}x,$$

subject to some known moment constraints or equations of moments.[1] Following Zellner and Highfield (1988), Ormoneit and White (1999), and Rockinger and Jondeau (2002), we consider only the arithmetic moments of the form

$$\int x^i p(x)\,\mathrm{d}x = \mu_i, \quad i = 0, 1, \ldots, k. \tag{1}$$

Extension to more general moments (e.g., the geometric moments, $E(\ln^i x)$ for $x > 0$) is straightforward (Soofi et al., 1995; Zellner and Tobias, 2001).

We use Lagrange's method to solve for the maxent density. The solution takes the form

$$p(x) = \exp\left(-\sum_{i=0}^{k} \lambda_i x^i\right), \tag{2}$$

where λ_i is the Lagrangian multiplier for the ith moment constraint. Since an analytical solution does not exist for $k \geqslant 2$, one must use a nonlinear optimization technique to solve for the maxent density. One way to solve the maxent problem is to transform the constrained optimization problem into an unconstrained optimization problem using the dual approach (Golan et al., 1996). Substituting Eq. (2) into the Lagrangian function and rearranging terms, we have the dual objective function for an unconstrained

[1] Mead and Papanicolaou (1984) give the necessary and sufficient condition for the moments that lead to a unique maxent density. We find that the sample moments of any finite sample satisfy this condition. The proof is available from the author upon request.

optimization problem

$$\Gamma = \ln Z + \sum_{i=1}^{k} \lambda_i \mu_i,$$

where $Z = e^{\lambda_0} = \int \exp(-\sum_{i=1}^{k} \lambda_i x^i)\,dx$.

Newton's method is used to solve for the Lagrange multiplier $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_k]'$ by iteratively updating

$$\boldsymbol{\lambda}_{(1)} = \boldsymbol{\lambda}_{(0)} - H^{-1} \frac{\partial \Gamma}{\partial \boldsymbol{\lambda}}, \tag{3}$$

where the gradient

$$\frac{\partial \Gamma}{\partial \lambda_i} = \mu_i - \frac{\int x^i \exp(-\sum_{i=1}^{k} \lambda_i x^i)\,dx}{\int \exp(-\sum_{i=1}^{k} \lambda_i x^i)\,dx} = \mu_i - \mu_i(\boldsymbol{\lambda}), \quad i = 1, 2, \ldots k$$

and the Hessian

$$H_{ij} = \frac{\partial^2 \Gamma}{\partial \lambda_i \partial \lambda_j} = \mu_{i+j}(\boldsymbol{\lambda}) - \mu_i(\boldsymbol{\lambda})\mu_j(\boldsymbol{\lambda}), \tag{4}$$

$$\mu_{i+j}(\boldsymbol{\lambda}) = \frac{\int x^{i+j} \exp(-\sum_{i=1}^{k} \lambda_i x^i)\,dx}{\int \exp(-\sum_{i=1}^{k} \lambda_i x^i)\,dx}, \quad i, j = 1, 2, \ldots k.$$

Since the Hessian matrix H is everywhere convex and therefore positive definite, there exists a unique solution. Mead and Papanicolaou (1984) show that the maxent estimates are consistent and efficient.

2. Sequential updating of the maxent density

In Bayesian analysis or information processing, it is known that the order in which information is incorporated into the learning process is irrelevant.[2] Hence instead of imposing all the moment constraints simultaneously, we can impose the moment constraints from lower to higher order and update the density estimates sequentially.

As shown in the previous subsection, solving for the maxent density subject to moment constraints $\boldsymbol{\mu}$ is equivalent to solving for the following system of equations:

$$\int x^i \exp\left(-\sum_{i=0}^{k} \lambda_i x^i\right) dx = \mu_i, \quad i = 0, 1, \ldots, k. \tag{5}$$

Since a unique solution exists, we can express $\boldsymbol{\mu}$ as a function of $\boldsymbol{\lambda}$. Denote $\boldsymbol{\mu} = f(\boldsymbol{\lambda})$, we know $f(\cdot)$ is a differentiable function since Eq. (5) is everywhere continuous and differentiable in $\boldsymbol{\lambda}$. By the Inverse Function Theorem, the inverse function of $\boldsymbol{\lambda} = f^{-1}(\boldsymbol{\mu}) = g(\boldsymbol{\mu})$ is also differentiable. Taking Taylor's expansion on $\boldsymbol{\lambda}$,

[2] See Zellner (1998) on the order invariance of maximum entropy procedures.

we obtain

$$\boldsymbol{\lambda} = g(\boldsymbol{\mu}_0 + \Delta\boldsymbol{\mu}) = g(\boldsymbol{\mu}_0) + g'(\boldsymbol{\mu}_0)\Delta\boldsymbol{\mu}.$$

This suggests that we can obtain the first-order approximation of $\boldsymbol{\lambda}$ corresponding to $\boldsymbol{\mu} = \boldsymbol{\mu}_0 + \Delta\boldsymbol{\mu}$, given $\boldsymbol{\lambda}_0 = g(\boldsymbol{\mu}_0)$ and $\Delta\boldsymbol{\mu}$.

For sufficiently small $\Delta\boldsymbol{\mu}$, one way to proceed is to use $\boldsymbol{\lambda}_0$ as initial values when we solve for $\boldsymbol{\lambda} = g(\boldsymbol{\mu})$ using Newton's method. If $\Delta\boldsymbol{\mu}$ is not small enough, we may not be able to obtain convergence for $\boldsymbol{\lambda} = g(\boldsymbol{\mu})$ using $\boldsymbol{\lambda}_0$ as initial values. In this case, we can divide $\Delta\boldsymbol{\mu}$ into M small segments such that $\Delta\boldsymbol{\mu} = \sum_{i=1}^{M} \Delta\boldsymbol{\mu}_i$ and solve for $\boldsymbol{\lambda}_m = g(\boldsymbol{\mu}_0 + \sum_{i=1}^{m} \Delta\boldsymbol{\mu}_i)$ using $\boldsymbol{\lambda}_{m-1}$ as initial values for $m = 1, \ldots, M$. However, this approach is rather inefficient because it involves a multi-dimension grid search for the k elements of $\boldsymbol{\mu}$. Instead, we can reduce the search to one dimension if we choose to impose the moment constraints sequentially.

Suppose for a given finite sample, we can solve for $\boldsymbol{\lambda}_k = g(\boldsymbol{\mu}_k)$, where $\boldsymbol{\mu}_k$ is the first k sample moments, using arbitrary initial values (usually a vector of zeros to avoid arithmetic overflow). Since higher moments are generally not independent of lower moments, the estimates from lower moments can serve as a proxy for the maxent density that is also subject to additional higher moments. Thus, if we fail to solve for $\boldsymbol{\lambda}_{k+1} = g(\boldsymbol{\mu}_{k+1})$ using arbitrary initial values, we can use $\boldsymbol{\lambda}'_{k+1} = [\boldsymbol{\lambda}_k; 0]$ as initial values. Note that the choice of zero as the initial value for λ_{k+1} is not simply for convenience, but is also consistent with the principle of maximum entropy. With only the first k moments incorporated into the estimates, λ_{k+1} for $p(x) = \exp(-\sum_{i=0}^{k+1} \lambda_i x^i)$ should be set to zero since no information is incorporated for the estimation of λ_{k+1}. In other words, if we do not use μ_{k+1} as side condition, the term x^{k+1} should not appear in the maxent density function. In this sense, zero is the 'most honest', or the 'most uninformative' guess for λ_{k+1}.

Corresponding to the 'most uninformative' guess $\lambda_{k+1} = 0$ is the predicted $(k+1)$th moment $v_{k+1} = \int x^{k+1} \exp(-\sum_{i=0}^{k} \lambda_k x^k)\, dx$, which is the unique maxent predicted value for μ_{k+1} based on the first k moments.[3] If v_{k+1} is close to μ_{k+1}, the difference $\Delta\boldsymbol{\mu}_{k+1}$ between the vector of actual moments $\boldsymbol{\mu}_{k+1}$ and $[\boldsymbol{\mu}_k; v_{k+1}]$ is small. Hence, if we use $\boldsymbol{\lambda}'_{k+1} = [\boldsymbol{\lambda}_k; 0]$ as initial values to solve for $\boldsymbol{\lambda}_{k+1} = g(\boldsymbol{\mu}_{k+1})$, the convergence can often be obtained in a few iterations. If we fail to reach the solution using $\boldsymbol{\lambda}'_{k+1}$ as initial values, we can divide the difference between v_{k+1} and μ_{k+1} into a few small segments and approach the solution using the above approach in multiple steps.

We note that the estimation of the maxent density becomes very sensitive to the choice of initial values as the number of moment constraints rises, partially because the Hessian matrix approaches singularity as its dimension increases. Fortunately, the difference between the predicted moment v_{k+1} based on the first k moments and the actual moment μ_{k+1} approaches zero as k increases. The higher k is, the closer is $p(x)$ to the underlying distribution, and subsequently the smaller the difference between μ_{k+1}

[3] Maximizing the entropy subject to the first k moments is equivalent to maximizing the entropy subject to the same k moments and the predicted $(k+1)$th moment v_{k+1}. Since v_{k+1} is a function of the first k moments, it is not binding when used together with the first k moments as side conditions. Therefore, the Lagrange multiplier λ_{k+1} for v_{k+1} is zero.

and the predicted moment v_{k+1}. Hence, the sequential method is especially useful when the number of moment constraints is large. On the other hand, sometimes we do not need to incorporate all the moment conditions. For example, the maxent density subject to the first moment is the exponential distribution of the form $p(x) = \exp(-\lambda_0 - \lambda_1 x)$ and the maxent density subject to the first two moments is the normal distribution of the form $p(x) = \exp(-\lambda_0 - \lambda_1 x - \lambda_2 x^2)$. So the first moment is the sufficient statistics for an exponential distribution and the first two moments are the sufficient statistics for a normal distribution. In this case, the difference between the predicted moment v_{k+1} and the actual moment μ_{k+1} can serve as a useful indicator to decide whether to impose more moment conditions.

3. Approximation of U.S. income distribution

In this section, we apply the sequential method to the approximation of the size distribution of U.S. family income. We run an experiment using U.S. family income data from the 1999 Current Population Survey (CPS) March Supplement. The data consist of 5,000 observations of family income drawn randomly from the 1999 March CPS. We fit the maxent density $p(x)=\exp(-\sum_{i=0}^{k} \lambda_i x^i)$ for k from 4 to 12 incremented by 2.[4] Newton's method with a vector of zeros as initial values fails to converge when the number of moment constraints k is larger than six, and we proceed with the sequential algorithm instead.

For the exponential family, the method of moments estimates are equivalent to maximum likelihood estimates.[5] Hence, we can use the log-likelihood ratio to test the function specification. Given $p(x_j) = \exp(-\sum_{i=0}^{k} \lambda_i x_j^i)$ for $j = 1, 2, \ldots, N$, the log-likelihood can be conveniently calculated as $L=\sum_{j=1}^{N} \ln p(x_j) = -N \sum_{i=0}^{k} \lambda_i \mu_i$, where μ_i is the ith sample moment. Since the maximized entropy subject to known moment constraints is $W = -\sum_{j=1}^{N} p(x_j) \ln p(x_j) = -\sum_{i=0}^{k} \lambda_i \mu_i$, the log-likelihood is equivalent to the maximized entropy multiplied by the number of observations. The first column of Table 1 lists the log-likelihood for the estimated maxent density and the second column reports the log-likelihood ratio of $p_{k+2}(x)=\exp(-\sum_{i=0}^{k+2} \lambda_i x^i)$ versus $p_k(x)=\exp(-\sum_{i=0}^{k} \lambda_i x^i)$. This log-likelihood ratio is asymptotically distributed as χ^2 with two degrees of freedom (critical value = 5.99 at 5% significance level). The log-likelihood ratio test favors the more general model $p_{k+2}(x)$ for our range of k.

Soofi et al. (1995) argue that the information discrepancy between two distributions can be measured in terms of their entropy difference. They define an index for comparing two distributions:

$$\mathrm{ID}(p, p^*) = 1 - \exp(-K(p : p^*)),$$

[4] Typically the income distribution is skewed with an extended right tail, which warrants including at least the first four moments in the estimation. Moreover, we should have even number of moment conditions to ensure that the density function integrates to unity.

[5] The maximum entropy method is equivalent to the ML approach where the likelihood is defined over the exponential distribution with k parameters. Golan et al. (1996) use a duality theorem to show this relationship.

Table 1
Specification and goodness-of-fit tests for estimated densities

	L (1)	LR (2)	ID (3)	KS (4)	AIC (5)	BIC (6)
$k=4$	2108	—	—	0.0300	0.8449	0.8569
$k=6$	2066	42.1	0.0084	0.0214	0.8288	0.8469
$k=8$	2048	18.6	0.0037	0.0174	0.8222	0.8463
$k=10$	2033	14.6	0.0029	0.0124	0.8172	0.8472
$k=12$	2020	13.3	0.0027	0.0065	0.8126	0.8487
Log-normal	2366	—	—	0.0507	0.9472	0.9532
Gamma	2115	—	—	0.0294	0.8466	0.8527

(1) Log-likelihood.
(2) Log-likelihood ratio test: $p_{k+2}(x)$ versus $p_k(x)$.
(3) Soofi (1995)'s ID index: $p_{k+2}(x)$ versus $p_k(x)$.
(4) Kolmogorov–Smirnov test.
(5) Akaike information criterion.
(6) Bayesian information criterion.

where $K(p:p^*)=\int p(x)\,(p(x)/p^*(x))\,\mathrm{d}x$ is the relative entropy or Kullback–Leibler distance, which is an information-theoretic measure of discrepancy between two distributions. The third column of Table 1 reports the ID indices between $p_{k+2}(x)=\exp(-\sum_{i=0}^{k+2}\lambda_i x^i)$ and $p_k(x)=\exp(-\sum_{i=0}^{k}\lambda_i x^i)$. We can see that the discrepancy decreases as more moment conditions enter the estimation. This suggests that as the number of moment conditions gets large, the information content of additional moment decreases.

We test the goodness-of-fit of the maxent density estimates using a two-sided Kolmogorov–Smirnov (KS) test. The fourth column of Table 1 reports the KS statistic of the estimated maxent density. The critical value of KS test at 5% significance level is 0.0192 for our sample. Thus, the KS test fails to reject the null hypothesis that our income sample is distributed according to $p_k(x)=\exp(-\sum_{i=0}^{k}\lambda_i x^i)$, for $k=8,10,12$.

To avoid overfitting, we calculate the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to check the balance between the accuracy of the estimation and the rule of parsimony. The results are reported in the fifth and sixth column of Table 1. The AIC test favors the model with 12 moment constraints. The BIC test, which has a greater complexity penalty, favors the model with the first eight moment constraints.

Lastly, we compare the maxent densities with two conventional income distributions. We fit a log-normal distribution and a gamma distribution to the income sample.[6] The relevant tests are reported in the last two columns of Table 1. Both of them fail the KS test and are outperformed by our preferred maxent densities in all the tests.

[6] The log-normal distribution and gamma distribution are in fact maxent densities subject to certain geometric moment constraints.

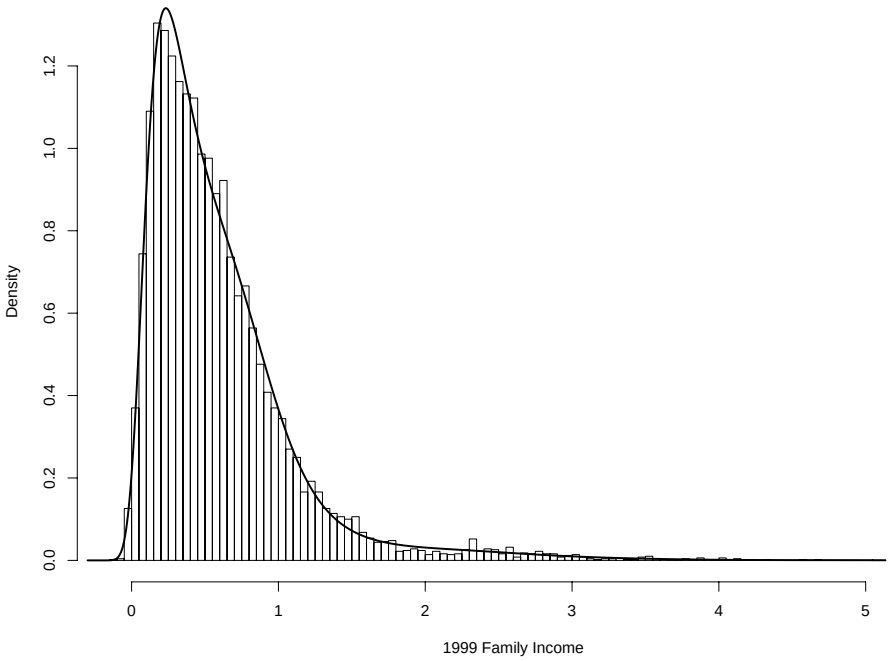

Fig. 1. Histogram and estimated maxent density for 1999 income, x-axis in \$100,000.

Fig. 1 reports the histogram of the income sample and the estimated maxent density with $k = 12$. The fitted density closely resembles the shape of the histogram of the sample. Although the domain over which the density is evaluated is considerably wider than the sample range in either end, the estimated density demonstrates good tail performance at both tails.

4. Summary

The maximum entropy approach is a flexible and powerful tool for density approximation. This paper proposes a sequential updating method for the maximum entropy density calculation. Instead of imposing the moment constraints simultaneously, this method incorporates the information contained in the moments into the estimation process from lower to higher moments sequentially. Consistent with the maximum entropy principle, we use the estimated coefficients based on lower moments as initial values to update the density estimates when additional higher-order moment constraints are imposed. The empirical applications on income distribution show the effectiveness of the proposed sequential updating method.

Acknowledgements

I am very grateful to Amos Golan, George Judge, Jeff LaFrance, Jeff Perloff, Stephen Stohs, Arnold Zellner and two anonymous referees for helpful suggestions and discussions.

References

Buchen, P., Kelly, M., 1996. The maximum entropy distribution of an asset inferred from option prices. Journal of Financial and Quantitative Analysis 31 (1), 143–159.

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

Hawkins, R., 1997. Maximum entropy and derivative securities. Advances in Econometrics 12, 277–301.

Jaynes, E.T., 1957. Information theory and statistical mechanics. Physics Review 106, 620–630.

Mead, L.R., Papanicolaou, N., 1984. Maximum entropy in the problem of moments. Journal of Mathematical Physics 25 (8), 2404–2417.

Ormoneit, D., White, H., 1999. An efficient algorithm to compute maximum entropy densities. Econometrics Reviews 18 (2), 127–140.

Rockinger, M., Jondeau, E., 2002. Entropy densities with an application to autoregressive conditional skewness and kurtosis. Journal of Econometrics 106, 119–142.

Soofi, E., Ebrahimi, N., Habibullah, M., 1995. Information distinguishability with application to analysis of failure data. Journal of Econometrics 90, 657–668.

Stutzer, M., 1996. A simple nonparametric approach to derivative security valuation. Journal of Finance 51 (5), 1633–1652.

Zellner, A., 1997. The Bayesian method of moments (BMOM): theory and applications. Advances in Econometrics 12, 85–105.

Zellner, A., 1998. On order invariance of maximum entropy procedures. Mimeo.

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

Zellner, A., Tobias, J., 2001. Further results on Bayesian method of moments analysis of the multiple regression model. International Economic Review 42 (1), 121–139.