Pub King

Risk Modeling in Agriculture: Retrospective and Prospective

Program proceedings for the annual meeting of the Technical Committee of S-232, held March 24-26, 1994, Gulf Shores State Park, Alabama.

Iowa State University
Ames, Iowa 50011-1070
August 1994

ENTROPY METHODS FOR RECOVERING INFORMATION FROM ECONOMIC MODELS

Douglas J. Miller*

Economics and Information Theory

Economists have long recognized the importance of information in behavioral models. Although information is potentially valuable if it allows agents to reduce uncertainty, it is also intangible and subjective. Consequently, information has been very difficult to incorporate in standard models of economic behavior. Prior to the flourish of interest in the economics of information that was led by developments in game theory and the publication of Akerlof's 'lemons' paper, Arrow concluded that information is "an economically interesting category of goods which has not hitherto been accorded much attention by economic theorists".

Works by Marschak, Arrow, and Theil are among the limited number of early attempts to use information in an economic context. Unlike more modern treatments of information, these research efforts referred to a body of literature known as information theory. Information theory is concerned with sending and receiving coded signals over noisy communication channels, and it is the foundation for research in computer science and electrical and electronic engineering. The economists applied information theory to the problem of measuring the amount of information contained in a given signal. The purpose of this paper is to introduce one measure of information, entropy, and to demonstrate methods that use entropy as a criterion for recovering information from data on observed economic behavior.

Entropy

To introduce the measure of information, consider a Bernoulli trial (e.g. the flip of a coin) that has success probability p. Given some p close to 1, we expect to observe a success and would be very surprised to observe a failure. If we receive a message or a signal that a success was observed, then the signal has very little informational content. Conversely, a signal reporting a failure would be very informative.

One way to convert the probabilities to a measure of 'informational content' is to use the minus-log transform. That is, define our measure of information in the signal as $H(p) = - 1\log(p) - 0\log(1-p) = - \log(p)$ for a success and $H(p) = - 0\log(p) - 1\log(1-p) = - \log(1-p)$ for a failure. Note that for a large value of p, H(p) is nearly zero for a success (we expected the message), and H(p) is infinitely large for a failure (we are 'shocked' to learn of a failure). Further, we can use $H(p) = - p\log(p) - (1-p)\log(1-p)$ as an ex ante measure of the expected information contained in the signal.

This functional form is the basis for a measure of information devised by Claude Shannon, who is often regarded as the founder of information theory. For an experiment that has K possible outcomes, Shannon's measure of uncertainty takes the following form:

---

*Formerly a Graduate Research Assistant at the University of California, Berkeley. Dr. Miller is currently an Assistant Professor at Iowa State University.

$$H(p) = -\sum_{i=1}^{K} p_i \cdot \log(p_i) \quad (1)$$

where p_i is the probability of observing the i^{th} outcome. Shannon named the measure entropy at the suggestion of John von Neumann, who recognized the similarity between Shannon's function and the measures of physical entropy used in statistical mechanics. Assuming 0log(0) = 0, H(**p**) takes on a maximum when **p** is the discrete uniform distribution, and it is minimized by a degenerate distribution (i.e. Pr[i] = 1 for some i). Thus, the distribution with 'maximum entropy' contains the most uncertainty about the outcome of the trial. When we know the outcome with certainty, the distribution has minimum entropy. Although all applications of H(**p**) in this paper use natural logarithms, we can normalize H to a unit scale by using base-K logarithms.

Cross-Entropy

A more general measure was later proposed by Kullback, and it is commonly known as Kullback-Leibler directed divergence, I-divergence, or cross-entropy:

$$I(p,q) = \sum_{i=1}^{K} p_i \cdot \log(p_i/q_i) \quad (2)$$

Here, **p** and **q** are different probability distributions over the same set of outcomes (support). For instance, **q** may be a set of prior probabilities, and **p** is the associated set of posterior probabilities. I(**p**,**q**) is then viewed as a measure of the information used to form the posterior from the prior distribution or as the amount of additional information reflected in the posterior. Alternately, **q** may represent our ex ante or prior beliefs, and **p** is the observed frequency distribution of outcomes from a number of trials. If **q** is a discrete uniform distribution, note that I(**p**,**q**) = - H(**p**) + log(K).

Although I(**p**,**q**) ≠ I(**q**,**p**) implies that I(**p**,**q**) is not a true distance function, we may still use it as a measure of the distance between **p** and **q** as we move with the flow of information (i.e. from prior to posterior). Although this flaw is not important in practice, we can use the joint entropy function, J(**p**,**q**) = I(**p**,**q**) + I(**q**,**p**), to form a 'true' distance measure if the commutative property is required.

The cross-entropy measure may be extended to more general probability measures by employing the Riemann-Stieltjes integral:

$$I(f,g) = \int \log[f(x)/g(x)]\,dF(x) \quad (3)$$

where the support of g contains the support of f. In this way, the entropy of continuous (Lesbesgue) or mixed distributions may be calculated. The entropy and related properties of various continuous distributions is summarized in a recent article by Maasoumi (1993). For convenience, the present discussion will be limited to discrete probability distributions.

Entropy and Information Recovery

To this point, we have used entropy as a diagnostic or descriptive tool. As with other measures like the sum of squared errors, we may also use entropy as a criterion for forming inferences about unknown parameters. Jaynes proposed the Method of Maximum Entropy as a feasible means of recovering a probability distribution from a set of moment conditions or other restrictions on the distribution. Given that observed economic data are often functions (e.g. averages or quantiles) of the underlying distribution of firms or consumers, entropy may be a useful criterion for solving a variety of problems in economics. The Method of Maximum Entropy is presented in the next section, and an extension proposed by Judge and Golan is discussed in the following section.

Method of Maximum Entropy

To motivate the Method of Maximum Entropy, consider Jaynes' dice problem. Suppose a six-sided die is rolled a large number of times, but we only observe the average of the outcomes. If the die is fair and follows the discrete uniform distribution, the Weak Law of Large Numbers implies that the observed average should converge (in probability) to 3.5. However, convergence to some other value (e.g. 4) is not consistent with the discrete uniform distribution (or any other distribution with a mean of 3.5). Given that there exist an infinite number of discrete distributions with a mean of $\mathbf{y} = 4$ and support $\mathbf{X} = [1, 2, 3, 4, 5, 6]$, the true distribution cannot be identified with certainty.

Jaynes noted that a conservative estimate of the unknown distribution would be the set of probabilities that satisfy the observed moment relations and are most uniform. Using Shannon's entropy as a measure of uniformity, Jaynes' Method of Maximum Entropy selects the distribution, $\mathbf{p}$, that solves the following problem:

$$Max\ H(p) = -\ p'log(p) \tag{4}$$

subject to:

$$y = Xp \quad and \quad p'1 = 1 \tag{5}$$

where $\mathbf{y}$ is a T-vector of observed moments and $\mathbf{X}$ is a (T x K) matrix containing the associated supports. The model restrictions in equation (5) are known as the consistency and additivity constraints, respectively. By including the observed information in the constraint set, the Maximum Entropy distribution "agrees with what is known, but expresses 'maximum uncertainty' with respect to all other matters" (Jaynes, 1968 p. 231). Note that this approach is a form of Generalized Method of Moments inference: the sample moments are equated with the population moments, and the equations are solved for the unknown probabilities under the entropy distance function.

The solution of the optimization problem has some special features to be noted. First, H is strictly concave on the interior of the additivity constraint set, $S^{K-1} = \{\mathbf{p}: \mathbf{p} \gg 0 \text{ and } p'1 = 1\}$, which is an open unit simplex. So, a solution exists if the intersection of the consistency constraint set and S^{K-1} is non-empty. Second, the first-order conditions (FOC) from the associated Lagrangian expression provide the following solutions:

$$p = exp(-X'\lambda) \,/\, [1'exp(-X'\lambda)] \qquad (6)$$

Clearly, the ME solutions are admissible because each $p_i \in (0,1)$ and the probabilities sum to 1. However, the FOC do not have a known closed-form solution, and the probabilities are expressed in terms of the unknown Lagrange multipliers on the consistency constraints, λ. Numerical optimization techniques must be used to solve the problem. Note that ignoring the consistency constraint yields the discrete uniform distribution ($p_i = 1/K$ for $i = 1,...,K$). Further, maximizing entropy without the additivity and consistency constraints results in $p_i = e^{-1}$ for all $i = 1,...,K$.

In the case of Jaynes' dice problem, we cannot recover the distribution by the traditional ML technique (using the multinomial distribution) because the frequency distribution over the outcomes is unknown. We often say that such problems are ill-posed because the number of unknowns (6) exceeds the number of identifying restrictions (2). However, we may proceed to solve the problem by the Method of Maximum Entropy. Given $y = 4$ and $\mathbf{X} = [1, 2, 3, 4, 5, 6]$, the numerical solution is:

$$p = [\ 0.103, 0.123, 0.146, 0.174, 0.207, 0.247\ ]$$

If the more extreme average of $y = 5.5$ is observed, then:

$$p = [\ 0.003, 0.009, 0.025, 0.075, 0.224, 0.664\ ]$$

Further, the discrete uniform distribution solves the inverse problem if $y = 3.5$, as expected.

Thus, Maximum Entropy provides one basis for recovering the probability distribution in a severely ill-posed problem. Aside from any intuitive appeal of Maximum Entropy, axioms and other principles of information theory have been used to justify Jaynes' approach. Given information in the form of constraints, Shore and Johnson (1980) prove that there is a unique, invariant distribution satisfying the constraints and other consistency axioms of information theory. They further demonstrate that this distribution may be recovered by maximizing entropy. Skilling and Csiszár employ alternate sets of axioms and achieve similar conclusions.

Note that Jaynes is implicitly assuming that the unknown probabilities are roughly uniform. The Method of Maximum Entropy may also be extended to include non-uniform (informative) prior information through the Method of Minimum Cross-Entropy. Given prior distribution $\mathbf{q}$, replace the objective function, $H(\mathbf{p})$, with $I(\mathbf{p},\mathbf{q})$ and minimizing this criterion subject to the same additivity and consistency constraints. Effectively, $\mathbf{p}$ is the least informative posterior distribution given our prior beliefs. Note that $I(\mathbf{p}, \mathbf{q}) = - H(\mathbf{p}) + \log(K)$ if $\mathbf{q}$ is a uniform density. So, maximizing entropy is equivalent to minimizing cross-entropy under uniform $\mathbf{q}$, and the Method of Maximum Entropy is a special case of the Method of Minimum Cross-Entropy.

Generalized Maximum Entropy

Critics of the Method of Maximum Entropy argue that the technique is not applicable when the moment conditions are observed with noise, $\mathbf{y} = \mathbf{Xp} + \mathbf{e}$, because the disturbance term, $\mathbf{e}$, is ignored. The disturbances may represent randomness in the behavior of economic agents

or measurement errors. To mitigate this problem, Judge and Golan (1992) proposed Generalized Maximum Entropy by rewriting the linear model as $\mathbf{y} = \mathbf{Xp} + \mathbf{e} = \mathbf{Xp} + \mathbf{Vw}$, or:

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ y_T \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ X_T \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \cdot \\ p_K \end{bmatrix} + \begin{bmatrix} W_1 & 0 & \cdot & 0 \\ 0 & W_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & W_T \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ v_T \end{bmatrix} \tag{7}$$

where $\mathbf{V}$ is a (T x JT) matrix containing the support of $\mathbf{w}$, a JT-vector of probabilities. So, the unknown disturbances are expressed as the expected values of the finite and discrete probability distribution, $\mathbf{w}$, and the reparameterized model conforms to the Maximum Entropy formalism.

Information about the unknown components of the general linear model may now be recovered by the Generalized Maximum Entropy. The problem is now to select $\mathbf{p}$ and $\mathbf{w}$ that maximize:

$$H(p,v) = - p'log(p) - w'log(w) \tag{8}$$

subject to:

$$y = Xp + Vw \tag{9}$$

$$\sum_{m=1}^{M} p_{km} = 1 \quad \forall\, k = 1,...,K \tag{10}$$

$$\sum_{j=1}^{J} w_{tj} = 1 \quad \forall\, t = 1,...,T \tag{11}$$

where equation (9) is the consistency constraint, and equations (10) and (11) provide the additivity constraints.

Judge and Golan also extended their reparameterization to incorporate prior information about the unknowns. Generalized Cross-Entropy minimizes the following objective function:

$$I(p,q,v,u) = \sum_{k=1}^{K} \sum_{m=1}^{M} p_{km}\log(p_{km}/q_{km}) + \sum_{t=1}^{T} \sum_{j=1}^{J} w_{tj}\log(w_{tj}/u_{tj}) \tag{12}$$

subject to the standard additivity and consistency constraints. Here, $\mathbf{q}$ and $\mathbf{u}$ are vectors of prior probabilities for the unknown probabilities and disturbances.

Recovering Markov Transition Probabilities

To demonstrate the potential of the entropy-based methods of recovering unknown probability distributions, consider data generated by a finite and discrete, first-order Markov process. A given process is characterized by the transition probability matrix, $\mathbf{P} = \{p_{ij}\}$, where each element represents the probability that an agent moves from state i to state j in one time period. For a collection of agents who behave according to the same process, the expected proportion of the group occupying state i at period t, x_{it}, may be computed as:

$$x_{it} = \sum_{j=1}^{K} x_{jt-1} p_{ji} \quad \forall\, i = 1,\ldots,K \tag{13}$$

where $\{x_{jt-1}\}$ is the frequency distribution of agents across the K states in the previous period, t-1. Further, equation (13) may be expressed in matrix form as $\mathbf{x_t} = \mathbf{x_{t-1}P}$, where $\mathbf{x_t}$ and $\mathbf{x_{t-1}}$ are K-vectors and **P** is the (K x K) Markov transition matrix. Finally, the Markov process is said to be stationary if **P** is time-invariant, and the observations from T periods may be assembled in matrix form. The transition relation may be rewritten as:

$$\begin{bmatrix} x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_T \end{bmatrix} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_{T-1} \end{bmatrix} P = ZP \Rightarrow \begin{bmatrix} x_{12} \\ \cdot \\ x_{1T} \\ x_{21} \\ \cdot \\ \cdot \\ \cdot \\ x_{KT} \end{bmatrix} = (I_K \otimes Z) \begin{bmatrix} p_{11} \\ \cdot \\ p_{K1} \\ p_{12} \\ \cdot \\ \cdot \\ \cdot \\ p_{KK} \end{bmatrix} \tag{14}$$

which takes the same form as the consistency constraint, $\mathbf{y} = \mathbf{Xp}$, where $\mathbf{p} = \text{vec}(\mathbf{P})$.

Given the linear model with noise, $\mathbf{y} = \mathbf{Xp} + \mathbf{e}$, researchers must recover **p** from observed values of **y** and **X**. One common approach is the least squares (LS) criterion. The properties of the LS estimator, $\mathbf{p} = (\mathbf{X'X})^{-1}\mathbf{X'y}$, are well-known and are discussed at length by Lee, Judge, and Takayama; Lee, Judge, and Zellner; and Madansky. Unfortunately, the LS estimates may be inadmissible for **p** (i.e. $p_i \notin [0,1]$ for some i = 1,...,K). Consequently, Lee, Judge, and Takayama note that an alternate approach is to solve the LS problem as a quadratic programming (QP) problem in which the boundary and additivity constraints are explicitly specified.

When additional equality or inequality constraints are imposed on the LS problem, the conditions of the Gauss-Markov Theorem are violated and the LS estimator may not be best linear unbiased. A popular alternative to the least squares criterion is the minimum absolute deviations (MAD) estimator, which was introduced as a generic curve-fitting technique (see Karst, Fisher). The advantages of MAD over LS estimators are presented in articles by Ashar and Wallace and by Bassett and Koenker. Finally, Kim and Schaible developed a generalized

form of the MAD estimator that simplifies the problem by reducing the number of variables. Their MOMAD estimator is named for the minimization criterion, the median of minimum absolute deviations.

To estimate **P** by these traditional methods, the problem must have more transition than states (i.e. T-1 > K) to be considered well-posed. Unfortunately, data are often limited so that there are more states than observed transitions conforming to a stationary Markov process. Hence, ill-posed problems may be common with Markov models. Given that the Method of Maximum Entropy is designed to estimate probabilities in ill-posed problems, we may recover an image of **P** by maximizing:

$$H(p) = -\sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij}\log(p_{ij}) \tag{15}$$

subject to:

$$x_t = x_{t-1}P \text{ for } t = 2,...,T \tag{16}$$

$$\sum_{j=1}^{K} p_{ij} = 1 \quad \forall\, i = 1,...,K \tag{17}$$

where equations (16) and (17) are the consistency and additivity constraints, respectively.

Transitions Observed without Noise

To examine the performance of the Method of Maximum Entropy in consider the following 4-state transition matrix devised by Lee, Judge, and Takayama:

$$P = \begin{bmatrix} 0.6 & 0.4 & 0 & 0 \\ 0.1 & 0.5 & 0.4 & 0 \\ 0 & 0.1 & 0.7 & 0.2 \\ 0 & 0 & 0.1 & 0.9 \end{bmatrix} \tag{18}$$

which was used to generate synthetic sample proportions for several transitions. The data may represent the aggregate behavior of an industry (e.g. size distribution of firms) or a collection of consumers (e.g. market shares of a good). Data for eleven periods (i.e. ten transitions) appear in the article by Kim and Schaible.

Suppose we only observed the last two transitions in the synthetic data set. From these eight sample proportions, the Method of Maximum Entropy recovers:

$$P_{ME} = \begin{bmatrix} 0.614 & 0.386 & 0 & 0 \\ 0.096 & 0.508 & 0.396 & 0 \\ 0 & 0.099 & 0.695 & 0.206 \\ 0 & 0 & 0.103 & 0.897 \end{bmatrix} \tag{19}$$

If we measure the accuracy of the estimators in terms of squared error loss (SEL), the Maximum Entropy probabilities are very close to correct with an SEL of just 5.48×10^{-4}.

Suppose we further restrict our information to the last transition, but know that the six "northeast" and "southwest" elements of P contain zeros (i.e. states 1 and 3, 1 and 4, and 2 and 4 do not communicate in one step). Given just four sample proportions, the Method of Maximum Entropy yields:

$$P = \begin{bmatrix} 0.434 & 0.566 & 0 & 0 \\ 0.148 & 0.366 & 0.486 & 0 \\ 0 & 0.125 & 0.332 & 0.544 \\ 0 & 0 & 0.298 & 0.702 \end{bmatrix} \tag{20}$$

with an SEL of 0.415.

Given that additional knowledge may exist, the Method of Maximum Entropy may be extended to incorporate informative prior information. Suppose we specify a matrix of modest prior beliefs, **Q**, that places a 50% chance of remaining in the same state and is uniform (or zero) elsewhere. Then, **Q** and the estimated transition matrix are:

$$Q = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.25 & 0.5 & 0.25 & 0 \\ 0 & 0.25 & 0.5 & 0.25 \\ 0 & 0 & 0.5 & 0.5 \end{bmatrix} \quad P = \begin{bmatrix} 0.457 & 0.543 & 0 & 0 \\ 0.141 & 0.574 & 0.345 & 0 \\ 0 & 0.084 & 0.467 & 0.449 \\ 0 & 0 & 0.243 & 0.757 \end{bmatrix} \tag{21}$$

and the SEL has declined from 0.415 to 0.215 by considering **Q**. Thus, we have enhanced the message contained in very limited data by incorporating prior information.

Finally, consider the relative performance of the Method of Maximum Entropy when the problem is well-posed. Kim and Schaible use several subsets of the transition data to compare

the accuracy of the QP, MAD, and MOMAD estimates. Using all of the data, their most accurate image of P:

$$P = \begin{bmatrix} 0.598 & 0.402 & 0 & 0 \\ 0.101 & 0.508 & 0.391 & 0 \\ 0 & 0.094 & 0.706 & 0.200 \\ 0 & 0.002 & 0.098 & 0.900 \end{bmatrix} \tag{22}$$

is recovered from the median-based MOMAD estimator with SEL = 2.34×10^{-5}.

From GAMS, the Maximum Entropy transition matrix is:

$$P = \begin{bmatrix} 0.603 & 0.397 & 0 & 0 \\ 0.099 & 0.501 & 0.400 & 0 \\ 0 & 0.100 & 0.700 & 0.200 \\ 0 & 0 & 0.100 & 0.900 \end{bmatrix} \tag{23}$$

with an SEL = 2.24×10^{-6}, which is less than one-tenth the MOMAD loss.

Kim and Schaible estimate **P** for several subsets of the data, which were also used here to derive the Maximum Entropy estimates. To conserve space, only the squared error loss of each estimate appears in Table 1 below. Note that the losses from the median techniques (MAD / MOMAD) are generally less than those generated by the QP estimator; Kim and Schaible used the greater accuracy of the median methods and the reduction in variables provided by MOMAD to promote their estimator. Further, Maximum Entropy losses are uniformly less than the losses resulting from the QP and MAD / MOMAD estimators. Thus, the entropy-based method of inference provides a substantial improvement over the standard techniques in this limited trial.

Table 1: Squared Error Loss of Competing Estimators			
Sample	QP	MAD/MOMAD	Maximum Entropy
9-18	0.077434	0.000234	0.0000224
10-18	0.067556	0.000306	0.000043
11-18	0.065884	0.001452	0.000289
12-18	0.325224	0.074104	0.000063
13-18	0.356956	0.362354	0.000603
14-18	0.63913	0.07633	0.0060
9-17	0.07201	0.000222	0.0000136
9-16	0.070144	0.000978	0.0000268
9-15	0.082204	0.000312	0.0000085
9-14	0.086366	0.030288	0.0000482
9-13	0.089232	0.07165	0.0000788
9-12	0.370968	0.000932	0.000261

Transitions Observed with Noise

To examine a Markov problem in which the transitions may be observed with noise, consider the study of the cigarette market completed by Telser. The market shares of the three major U. S. cigarette brands (Camel, Lucky Strike, and Chesterfield) were collected for 1925-43. Telser then estimated the Markov transition probabilities in order to measure brand loyalty patterns among U.S. smokers.

Assuming all smokers behave according to a stationary, first-order Markov process, the QP estimate of the 3-state Markov transition matrix is:

$$P = \begin{bmatrix} 0.6686 & 0.1423 & 0.1891 \\ 0 & 0.8683 & 0.1317 \\ 0.4019 & 0 & 0.5981 \end{bmatrix} \tag{24}$$

Lee, Judge, and Zellner present estimates from the other traditional techniques, but the results are similar to the QP estimate and are not included here. Given the relatively large probabilities along the diagonal of the transition matrix, Telser concludes that there was a considerable degree of brand loyalty among cigarette consumers within the sample period.

The Method of Maximum Entropy was also employed, but the numerical optimization algorithm declared the inverse problem to be infeasible. The difficulty may be due to sampling errors in Telser's data. The consistency constraints do not account for the disturbances, and the entropy method must accept all of the variability in y as signal variation. Consequently, the algorithm is not able to find a candidate p that represents the systematic and noise components as purely signal variation. Thus, the consistency constraints are not satisfied because they are misspecified.

Alternately, consider the image of P recovered by the generalized method of Judge and Golan. Here, the support used for each of the disturbances is centered about zero and is $\mathbf{V}_j$ = [-0.5, -0.4,..., 0,..., 0.4, 0.5] for j = 1,...,T. Further, suppose moderately informative prior beliefs about the transition probabilities, Q, and uniform prior beliefs about the distribution of the errors are maintained. The resulting image recovered by Generalized Maximum Entropy is:

$$P = \begin{bmatrix} 0.664 & 0.143 & 0.194 \\ 0.035 & 0.786 & 0.180 \\ 0.363 & 0.102 & 0.535 \end{bmatrix} \tag{25}$$

which is comparable to the QP estimate.

The lengthy sample extends over a very turbulent economic period, and the stationarity assumption required for the standard estimation techniques is difficult to justify. Because the entropy methods are not confined to well-posed problems, it is possible to estimate P for smaller and perhaps more homogeneous samples. If we believe the behavior of cigarette smokers varied over the periods of expansion, depression, and war, the sample may be divided into three distinct subsamples: 1925-29, 1929-39, and 1939-43. The resulting images of P are:

$$P_{25-29} = \begin{bmatrix} 0.720 & 0.020 & 0.260 \\ 0.074 & 0.826 & 0.099 \\ 0.252 & 0.228 & 0.519 \end{bmatrix} \quad P_{29-39} = \begin{bmatrix} 0.646 & 0.210 & 0.144 \\ 0.028 & 0.783 & 0.189 \\ 0.405 & 0.009 & 0.586 \end{bmatrix}$$

$$P_{39-43} = \begin{bmatrix} 0.517 & 0.316 & 0.166 \\ 0.272 & 0.378 & 0.350 \\ 0.279 & 0.351 & 0.371 \end{bmatrix} \tag{26}$$

given the same prior beliefs about the probabilities and the error distributions.

Summary and Conclusions

Based on the limited results presented above, entropy appears to be a very useful diagnostic tool for economic analysis as well as a criterion for forming inferences about unknown probability distributions. Although Jaynes devised the Maximum Entropy formalism to solve ill-posed problems for an unknown probability distribution, the method may be extended to recover several distributions from data observed with noise. Judge and Golan also demonstrate that Generalized Maximum Entropy may be used to provide reasonable inferences in problems with real-valued unknowns. Specific applications include the general linear model, systems of simultaneous linear equations, non-stationary stochastic processes, and qualitative choice models.

References

Anderson, T. W., and L. A. Goodman. 1957. "Statistical Inference about Markov Chains." *Ann. Math. Stat.* 28: 89-110.

Arrow, K. J. 1984. *The Economics of Information.* Vol. 4 of *Collected Papers of Kenneth J. Arrow.* Cambridge: Harvard University Press.

Ashar, V. G., and T. D. Wallace. 1963. "A Sampling Study of Minimum Absolute Deviations Estimators." *Operations Research* 11: 747-58.

Bassett, G., and R. Koenker. 1978. "Asymptotic Theory of Least Absolute Error Regression." *JASA* 73: 618-22.

Brooke, A., D. Kendrick, and A. Meeraus. 1992. *GAMS: A User's Guide, Release 2.25.* South San Francisco: Scientific Press.

Csiszár, I. 1991. "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems." *Ann. Stat.* 19: 2032-66.

Donoho, D. L, I. M. Johnstone, J. C. Hoch, and A. S. Stern. 1992. "Maximum Entropy and the Near Black Object." *J. Roy. Stat. Soc., Ser. B.* 54: 41-81.

Fisher, W. D. 1961. "A Note on Curve Fitting with Minimum Deviations by Linear Programming." *JASA* 56: 359-63.

Good, I. J. "1963. Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables." *Ann. Math. Stat.* 34: 911-34.

Jaynes, E. T. 1957a. "Information Theory and Statistical Mechanics, I." *Physics Review* 106: 620-30.

_____. 1957b. "Information Theory and Statistical Mechanics, II." *Physics Review* 108: 171-190.

_____. 1968. "Prior Probabilities." *IEEE Transactions Syst. Sci. Cybern.* SSC-4: 227-41.

Judge, G. G., and A. Golan. 1992. "Recovering Information in the Case of Ill-posed Inverse Problems with Noise." (Unpublished manuscript), University of California, Berkeley.

Karlin, S., and H. M. Taylor. 1975. *A First Course in Stochastic Processes,* 2nd ed. San Diego: Academic Press.

Karst, O. J. 1958. "Linear Curve Fitting Using Least Deviations." *JASA* 53: 118-32.

Kim, C. S., and G. Schaible. "1988. Estimation of Transition Probabilities Using Median Absolute Deviations." *J. Agr. Econ. Research.* 40: 12-19.

Kullback, J. 1959. *Information Theory and Statistics*. New York: Wiley.

Laffont, J.-J. 1989. *The Economics of Uncertainty and Information*. Cambridge: MIT Press.

Lee, T. C., G. G. Judge, and T. Takayama. 1965. "On Estimating the Transition Probabilities of a Markov Process." *J. Farm Econ.* 47: 742-62.

Lee, T. C., G. G. Judge, and A. Zellner. 1977. *Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data*. Amsterdam: North-Holland.

Maasoumi, E. 1993. "A Compendium to Information Theory in Economic and Econometrics." *Econometric Reviews* 12: 137-81.

Madansky, A. 1959. "Least Squares Estimation in Finite Markov Processes." *Psychometrika* 24: 137-44.

Marschak, J. 1959. "Remarks on the Economics of Information." in *Contributions to Scientific Research in Management*. Los Angeles: Western Data Processing Center, University of California, pp. 79-98.

O'Sullivan, F. 1986. "A Statistical Perspective on Ill-posed Inverse Problems." *Stat. Science* 1: 502-27.

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27: 379-423.

Shore, J. E., and R. W. Johnson. 1980. "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy." *IEEE Transactions on Information Theory* IT-26: 26-37.

_____. 1981. "Properties of Cross-Entropy Minimization." *IEEE Transactions on Information Theory* IT-27: 472-82.

Skilling, J. 1988. "The axioms of maximum entropy." in *Maximum Entropy and Bayesian Methods in Science and Engineering*. Amsterdam, Kluwer, 1: 173-87.

Telser, L. G. 1963. "Least-Squares Estimates of Transition Probabilities." Chap. 11 in C. F. Christ, *et al*, eds., *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*. Stanford: Stanford University Press.

Theil, H. 1967. *Economics and Information Theory*. Amsterdam: North-Holland.