COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station

New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 868

KNIGHTIAN DECISION THEORY AND ECONOMETRIC INFERENCE

Truman F. Bewley

March 1988

KNIGHTIAN DECISION THEORY AND ECONOMETRIC INFERENCE

Truman F. Bewley

March 1988

*I have greatly benefited from discussions with Professor Burt Singer. I have also benefited from the advice of Professors John Hartigan and James Berger. This work has been suported by National Science Foundation Grant SES-8605046.

1. Introduction

In previous papers (1986, 1987), I have proposed a Knightian theory of decision in which decision makers use many subjective probability distributions to compare alternatives. The multiplicity of the set of subjective distributions arises from ignorance of the true probabilities and aversion to uncertainty. Following Frank Knight (1921), I use the word "uncertainty" to mean random variation according to an unknown probability law. "Risk" means random variation according to a known law. Uncertainty aversion is distinct from risk aversion. An increase in uncertainty aversion increases the multiplicity of the subjective distributions. An uncertainty neutral decision maker has only one subjective distribution and so is Bayesian.

In this paper, I attempt to reconcile the apparent definiteness of econometric practice with the vagueness of subjective probabilities assumed in Knightian decision theory. I argue that some standard uses of classical inference are Knightian in spirit, even though the formal justification of classical methods uses the frequentist notion of probability. Classical confidence regions may be viewed as defining sets of posterior means corresponding to a standardized set of prior distributions. Tests of the null hypothesis that a parameter equals a particular value may be viewed as determining whether it is rational, from a Knightian point of view, to act as if the null hypothesis were true. This interpretation of the tests seems to correspond fairly well to practice and to the informal story told by classical statisticians. Hence, one could argue that to this extent classical statisticians act unconsciously as Knightian decision makers. If one accepts this argument, then it is of interest to know what level of uncertainty aversion corresponds to the popular 5% significance level. The

definition of standardized sets of prior distributions involves a standardized and measurable form of uncertainty aversion. The equivalence of classical confidence regions and sets of posterior means establishes a correspondence between confidence levels and levels of risk aversion. In examples, the levels of risk aversion corresponding to the 95% confidence level are quite high. This fact is indirect evidence that uncertainty aversion is a real and important aspect of reality.

The Knightian uncertainty associated with parameter estimation would tend to disappear as data was accumulated. This fact seems to imply that econometric research would eventually remove nearly all Knightian uncertainty about the economic environment. But this would be so only if the environment were governed by probability laws which could be inferred from data. I argue that not every stochastic process has such a law. I call such a law discoverable and give a tentative definition of discoverability. I prove that not every process has a discoverable law and also that every discoverable law may be learned. I also discuss how one might test for discoverability. Any failure of a model to pass a test for structural stability is evidence in favor of lack of discoverability. I believe that lack of discoverability is not an unlikely hypothesis to apply to many economic time series.

The paper is organized as follows. In the next section, I review Knightian decision theory briefly. In Section 3, I define a standardized and measurable form of uncertainty aversion. Section 4 discusses, from a Knightian viewpoint, confidence intervals and related tests of hypotheses. The ideas of Section 4 are applied to the normal linear regression model in Section 5. Related matters are discussed in Sections 6 and 7. Discover-

ability is discussed in Sections 8 and 9. In Section 10, I interpret the Kalman filter model as expressing Knightian uncertainty about the evolution of parameters in a linear regression model. In an appendix, I address the issue of how one can speak of Knightian or Bayesian subjective probabilities over unobservable distributional hypotheses. Subjective probabilities are inferred from preferences over bets, but one cannot bet on an unobservable event, such as the truth of some distributional hypothesis. I show that if one makes a certain simple assumption, then distributions over hypotheses can be derived from preferences over bets on only observable events.

Edward Leamer (1987) has argued that probability intervals should be used to express uncertainty about both prior and posterior probabilities. He emphasizes the connection between this Knightian view and the analysis of the sensitivity of posterior to prior distributions. Peter Walley (1984) is writing or has written an unpublished book which advocates the use of multiple subjective probabilities as a basis for statistical inference and decision theory. I have been able to obtain only the introduction.

This paper of mine should not be understood as advocating that Knightian decision theory be the foundation of statistical inference. I have no opinion on that matter. My purposes are to describe how Knightian uncertainty could be quantified and to imagine how it could persist in a world we think of as governed by definite stochastic processes.

2. Review

I describe briefly the basic ideas and notation of the Knightian decision theory presented in Bewley (1986, 1987). The starting point is a relation of strict preference, $\succ$, defined over a topological vector space, X, of lotteries. The space X consists of measurable real-valued functions on a state space S with σ-field $\mathcal{S}$. The topology of X is locally convex and the space, X^*, of continuous linear functionals on X may be identified with the signed measures on $\mathcal{S}$. The ordering $\succ$ is transitive, but may not be the strict preference posterior of a complete ordering. (An ordering, $\succsim$, is complete if and only if for every x and y in X, either $x \succsim y$ or $y \succsim x$.) The ordering $\succ$ is continuous and satisfies a substitution assumption. Under these assumptions, $\succ$ may be characterized by a convex and closed set, $\Pi(\succ)$, of measures on $\mathcal{S}$ as follows: $x \succ y$ if and only if $\int(x(s) - y(s))\pi(ds) > 0$, for all $\pi \in \Pi(\succ)$. Depending on context, $\Pi(\succ)$ will be either a cone of measures or a set of probability measures. The characterization is proved by noting that $C(\succ) = \{x \in X | x \succ 0\}$ is a convex cone containing the positive cone of X. The set $\Pi(\succ)$ is dual to $C(\succ)$.

An additional assumption guarantees that if the probability of an event is known objectively, then the subjective probabilities of the event all equal the objective one. Thus, the von Neumann-Morgenstern theory of choice under risk is not contradicted. Incompleteness of preferences is what distinguishes behavior toward uncertainty from behavior toward risk.

The preference ordering $\succ$ is said to be <u>uncertainty averse</u> if and only if $\Pi(\succ)$ intersects more than one ray through the origin in X^*. The ordering is <u>uncertainty neutral</u> if and only if $\Pi(\succ)$ is contained in a

single ray of X^* , which in turn is true if and only if $>$ is the strict preference part of a complete preference ordering. Thus, uncertainty neutrality is equivalent to being Bayesian.

The term "uncertainty aversion" stems from an additional hypothesis called the inertia assumption. According to this assumption, it is possible to define a status quo, and the status quo is abandoned only in favor of a point strictly preferred to it. Thus, increased uncertainty aversion implies increased conservatism.

If $B \subset X$ is a choice set, then $\underline{b} \in B$ is said to be undominated or maximal if there is no $b \in B$ such that $b \succ \underline{b}$. A basic theorem, valid for convex B under quite general conditions, is that $\underline{b}$ is maximal if and only if there is $\pi \in \Pi(\succsim)$ such that $\underline{b}$ solves the problem $\max_{b \in B} \int b(s)\pi(ds)$. Thus, $\underline{b}$ is maximal if and only if it is optimal with respect to some π . A Knightian decision maker acts as a Bayesian one, except when comparing a new alternative to the status quo. The theorem is proved by separating B from $\underline{b} + C(\succsim)$.

3. A Measurable Form of Uncertainty Aversion

There seems to be no useful measure of uncertainty aversion which may be applied to an arbitrary Knightian preference relation. One cannot even compare the uncertainty aversion of two preference orderings, $\succsim$ and $\succsim'$, over the same set of lotteries if one of the cones $C(\succsim)$ or $C(\succsim')$ does not contain the other. However, one may obtain a measure of uncertainty aversion inherent in an individual by considering the size of the set of probabilities he attaches to two disjoint events about whose likelihood he is equally and totally ignorant. The symmetry of the ignorance can be

interpreted as meaning that the probabilities attached to one event are symmetric about 1/2. Letting these probabilities be the interval $\left[\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon\right]$, the number $a = \left(\frac{1}{2} - \varepsilon\right)^{-1}\left(\frac{1}{2} + \varepsilon\right) - 1$ is an index of the individual's inherent uncertainty aversion.

Consider now the set of probabilities attached by the same individual to N events about whose likelihood he is equally and totally ignorant. If the individual is consistent, the set of probabilities attached to one of any two events, conditional on their union, should be of the form $\left[\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon\right]$. This restriction is not enough to define the set of probabilities over the N events. However, there is a unique largest set of probabilities satisfying those conditions, namely, $\Pi_{N,a} = \{(\pi_1, \ldots, \pi_N) | \pi_n \geq 0$, for all n, $\Sigma\pi_n = 1$ and $\pi_k \leq (1+a)\pi_n$, for all k and n$\}$, where π_n is the probability of the n^{th} event. This set may be thought of as a standardized set of subjective distributions over N alternatives expressing equal and total ignorance and uncertainty aversion of degree a.

The definition of such a standardized set of distributions may be extended to continua of events, provided one takes as given some measure expressing the equivalence of sets about whose likelihood one is equally ignorant. Let λ be such a measure defined on the set of events S. The measure λ might be improper in the sense that $\lambda(S)$ is infinite. For instance, if S were the real line, λ might be Lebesgue measure. The set of measures expressing equal ignorance relative to λ and uncertainty aversion of degree a is $\Pi_{\lambda,a} = \{\pi | \pi$ is a measure on S and if, for A and B in S, $A \cap B = \phi$ and $0 < \lambda(A) = \lambda(B) < \infty$, then $\pi(A) \leq (1+a)\pi(B)\}$. If λ is a non-atomic σ-finite measure on (S,S), then $\Pi_{\lambda,a}$ may be taken to be $\{\pi | \pi$ is the indefinite integral with respect to λ of a measurable function

$f : S \to R$ such that $1 \leq f(s) \leq 1+a$, for all $s \in S$) .

When applied as prior distributions over parameters, the diffuse or uninformative distributions of Bayesian statistical theory seem to be intended to express scientific detachment. Thus, if λ is diffuse, the set $\Pi_{\lambda,a}$ expresses scientific detachment together with conservatism of degree a .

4. Knightian Statistical Inference

Standardized sets of prior distributions may be used to obtain Knightian analogues of certain classical confidence regions and of associated tests of hypotheses. These Knightian confidence regions and tests have the advantage that they have a decision theoretic interpretation. Let y be a random variable with a probability distribution depending on a vector, s , of unknown parameters belonging to a set S . The values of y may be finite dimensional vectors. Suppose that one can define a diffuse prior distribution, λ over S , so that $\Pi_{\lambda,a}$ is a standardized set of priors for s . Finally, let $\underline{y} = (y_1, \ldots, y_N)$ be any random sample from the distribution for y . Given $\underline{y}$, there corresponds to each π in $\Pi_{\lambda,a}$ a posterior probability distribution for s and hence a posterior distribution for y , call it $p_\pi(dy|\underline{y})$. Let $P_{\lambda,a}(y|\underline{y}) = \{p_\pi(dy|\underline{y}) : \pi \in \Pi_{\lambda,a}\}$. The set of posterior means, $M_{\lambda,a}(y|\underline{y}) = \{\int y(s)p(ds) : p \in P_{\lambda,a}(y|\underline{y})\}$ is the Knightian-analogue of the classical confidence interval for the mean of y . Notice that y need not be normally distributed in order for $M_{\lambda,a}(y|\underline{y})$ to be well-defined. $M_{\lambda,a}(y|\underline{y})$ is well-defined as long as the posterior means of y exist. Notice that the regions $M_{\lambda,a}(y|\underline{y})$ are indexed by the degree of uncertainty aversion rather than the confidence

level.

The decision theoretic interpretation of $M_{\lambda,a}(y|\underline{y})$ stems from the fact mentioned in Section 2 that decisions are undominated in the Knightian sense if and only if they are optimal with respect to some subjective distribution. Suppose that after observing $\underline{y} = (y_1, \ldots, y_N)$, a decision maker with priors $\Pi_{\lambda,a}$ must make choices whose payoffs in utility depend linearly on a future random sample from the distribution of y, say $y_{N+1}, \ldots, y_{N+K}$. Then, a maximal decision will be optimal with respect to some one of the posterior distributions for y. Since the dependence of payoffs on the future sample is linear, the decision maker need know only the mean of the posterior distribution. Thus, $M_{\lambda,a}(y|\underline{y})$ is the set of means it would be rational to use to evaluate alternative choices.

This assertion may be demonstrated more rigorously as follows. The payoff of a choice is a linear function of the future sample, say $b_0 + b_1 \cdot y_{N+1} + \ldots + b_k \cdot y_{N+K}$, where b_0 is a number and $b_1, \ldots, b_K$ are vectors of the same dimension as y. Represent the choice by the vector $b = (b_0, b_1, \ldots, b_K)$ and let B be the set of choices. Corresponding to any posterior distribution $p \in P_{\lambda,a}(y|\underline{y})$, there is a linear functional, F_p, on B defined by $F_p(b) = b_0 + (\int y(s)p(ds)) \cdot \sum_{k=1}^{K} b_k$. A choice $\underline{b}$ in B is maximal if and only if there is no b in B is such that $F_p(b) > F_p(\underline{b})$, for all p in $P_{\lambda,a}(y|\underline{y})$. Clearly, if for some p in $P_{\lambda,a}(y|\underline{y})$, $F_p(\underline{b}) = \max_{b \in B} F_p(b)$, then $\underline{b}$ is maximal. If B is convex, it follows from Minkowski's separation theorem that $\underline{b}$ in B is maximal only if there is $p \in P_{\lambda,a}(y|\underline{y})$ such that $F_p(\underline{b}) = \max_{b \in B} F_p(b)$. Since the linear functional F_p depends on p only through its mean $\int y(s)p(ds)$, it follows that the set $M_{\lambda,a}(y|\underline{y})$ is the set of means which may rationally be

used to evaluate alternatives. Loosely speaking, it would be rational to act as if the true mean of y were any point in $M_{\lambda,a}(y|\underline{y})$.

Of course, if payoffs depended non-linearly on future observations of y , then the decision maker would need to know more than simply the mean of a posterior distribution in order to evaluate alternatives.

The regions $M_{\lambda,a}(y|\underline{y})$ may be used to test hypotheses about the mean of y . Define the hypothesis to be a subset, H , of the vector space to which y belongs. Say that H is accepted if and only if $H \cap M_{\lambda,a}(y|\underline{y}) \neq \phi$. The decision theoretic interpretation is that in the context of the decision problem previously defined, it is rational to act as if the true mean of y satisfied the hypothesis.

This interpretation of hypothesis testing applies also to tests of the point hypothesis that the mean of y is a particular value, say $\bar{y}$. This hypothesis is accepted if and only if $\bar{y} \in M_{\lambda,a}(y|\underline{y})$. Acceptance of the hypothesis implies that it is rational for the decision maker to act as if the mean of y were $\bar{y}$.

The above discussion can be extended to statistical inference about linear regression parameters. Suppose that it is known that $y = \beta_0 + \sum_{k=1}^{K} \beta_k x_k + \varepsilon$, where the β_k are unknown constant vectors of the same dimension as y , the x_k are predetermined numbers or exogenous random variables and ε is an error of mean zero. Suppose that a diffuse prior distribution, λ , is given over the β_k's and the parameters of the distribution of ε . Finally, suppose that past observations of the x_k's and the y's are given. Call the vector of these observations $(\underline{x},\underline{y})$. Then, to each prior in $\Pi_{\lambda,a}$, there corresponds a posterior mean of the vector $\beta = (\beta_0, \ldots, \beta_K)$. Let $M_{\lambda,a}(\beta|\underline{x},\underline{y})$ be the set of these posterior

means, one for each prior in $\Pi_{\lambda,a}$. $M_{\lambda,a}(\beta|\underset{\sim}{x},\underset{\sim}{y})$ is the Knightian confidence region for β . Letting a hypothesis be a subset, H , of the space of all possible vectors, β , H is accepted if and only if $H \cap M_{\lambda,a}(\beta|\underset{\sim}{x},\underset{\sim}{y}) \neq \phi$.

A decision theoretic interpretation of this form of inference may be given by assuming again that a decision maker with priors $\Pi_{\lambda,a}$ faces a decision problem with payoffs depending linearly on future values of y . Suppose also that he can forecast or predetermine future values of the x_k . Then, he acts rationally if he uses any point β in $M_{\lambda,a}(\beta|\underset{\sim}{x},\underset{\sim}{y})$ to evaluate his alternatives. The point hypothesis $\beta = \bar{\beta}$ is accepted if and only if it is rational for him to act as if β were the true mean. Notice that the Knightian test of the hypothesis $\beta = \bar{\beta}$ does not require that a positive prior probability be attached to the single point, $\bar{\beta}$, as must be done in Bayesian tests of this hypothesis.

5. The Normal Linear Regression Model

The Knightian confidence regions may be said to be analogous to the classical regions because the two coincide in the case of the normal linear regression model. Also, Knightian tests of linear hypotheses on the model are equivalent to the classical tests.

Represent the normal linear regression model by the equations $y_t = \beta_0 + \sum_{k=1}^{K} x_{tk}\beta_k + \varepsilon_t$, for $t = 1, \ldots, T$, where the y_t and x_{tk} are observed numbers, the β_k are unknown constants, and the ε_t are mutually independent normally distributed random variables with mean zero and variance σ^2 . In vector form, the model is $y = X\beta + \varepsilon$, where $y = (y_1, \ldots, y_T)'$, $\beta = (\beta_0, \ldots, \beta_K)'$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)'$ and X is

the obvious $T \times (K+1)$ matrix of ones and x_{tk}'s . Throughout what follows, it will be assumed that X has rank $K+1$.

The choice of a diffuse distribution is a controversial subject.[1] It can by no means be said that there is any obviously standard diffuse prior on the parameter vectors (β,σ) . Perhaps the most attractive property that such a prior should have is to be invariant to changes in the units in which the data are measured. Similarly, the distribution should remain invariant if linear combinations of the x_{tk}'s are added to the y_t . Suppose that the x_{tk} are replaced by $\hat{x}_{tk} + a_{xk} + b_{xk}x_{tk}$, where a_{xk} and b_{xk} are constants. Suppose also that the y_t are replaced by $\hat{y}_t = a_y + by_t + \sum_{k=1}^{K} c_k x_{tk}$, where a_y , b and the c_k are constants. Then, $\hat{y}_t = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k \hat{x}_{tk} + \hat{\varepsilon}_t$, where $\hat{\varepsilon}_t$ is normally distributed with mean zero and variance $\hat{\sigma}^2 = b^2\sigma^2$ and where the vector $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_K)'$ is of the form $\hat{\beta} = d + D\beta$, d being a vector of constants and D being an invertible matrix of constants. If $f(\beta,\sigma)$ is the prior density, then the transformed density of $(\hat{\beta},\hat{\sigma})$ is $F(\hat{\beta},\hat{\sigma}) = b^{-1}|D|^{-1}f(D^{-1}(\hat{\beta}-d), b^{-1}\hat{\sigma})$. The desired invariance property is that $F(\hat{\beta},\hat{\sigma}) = Cf(\hat{\beta},\hat{\sigma})$, for some constant C . Since f is improper, C need not equal one. The invariance property implies that $f(\beta,\sigma) = A\sigma^n$, for some constants A and n .

In order to fix the exponent n , one may require that the prior distribution of y transform correctly under changes in the scale of y . The prior density of y , given the prior density $A\sigma^n$ over (β,σ) is $p(y) = A(2\pi)^{-T/2}\int\int\sigma^{n-T}\exp\left[-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right]d\beta d\sigma$. If y is replaced by by , where b is a non-zero number, then the prior density $p(y)$ trans-

[1] Zellner (1971) has a good discussion of diffuse priors in the appendix of Chapter II.

forms to $P(by) = b^{-T}p(y)$. The required invariance property is that $P(by) = p(by)$. It is not hard to calculate that $p(y) = \hat{A}s^{-(T-K-n-2)}$ where $\hat{A}$ is a constant and $s^2 = (T-K-1)^{-1}(y - (X'X)^{-1}X'y)'(y - (X'X)^{-1}X'y)$ is the usual unbiased estimate of σ^2 . One sees that if $n = -(K+2)$, then $p(y) = \hat{A}s^{-T}$ and so $P(by) = p(by)$, as desired.

Let λ be the prior distribution over (β,σ) with density $\sigma^{-(K+2)}$ As in the previous section, let $M_{\lambda,a}(\beta|X,y)$ be the set of posterior means of β given (X,y) and calculated using priors in $\Pi_{\lambda,a}$.

The theorem below expresses the equivalence of $M_{\lambda,a}(\beta|X,y)$ with the classical confidence region for β . In the statement of the theorem, $b = (X'X)^{-1}X'y$ is the least squares estimate of β and s^2 is as before. Also, F_T denotes the cumulative distribution function of the student t distribution with T degrees of freedom and Γ denotes the gamma function.

5.1) Theorem. $M_{\lambda,a}(\beta|X,y) = \{\beta|(\beta-b)'X'X(\beta-b) \leq \gamma^2 s^2\}$, where $\xi = \sqrt{T/(T-K-1)}\ \gamma$ is the unique solution of the equation

$$\xi = \frac{a}{a+1}\left[\xi F_T(\xi) + \frac{T\Gamma\left[\frac{T+1}{2}\right]}{(T-1)\Gamma\left[\frac{T}{2}\right]\sqrt{T\pi}}\left(1 + \frac{\xi^2}{T}\right)^{-(T-1)/2}\right] .$$

The classical confidence region for β of confidence level 100α would be $\{\beta|(\beta-b)'X'X(\beta-b) \leq \delta s^2(1+K)\}$, where $F^{K+1}_{T-K-1}(\delta) = \alpha$ and F^{K+1}_{T-K-1} is the cumulative distribution function of the Snedecor F distribution with $K+1$ and $T-K-1$ degrees of freedom. Clearly, for any $\delta > 0$ there exists an α such that $F^{K+1}_{T-K-1}(\delta) = \alpha$. Thus, $M_{\lambda,a}(\beta|X,y)$ is a classical confidence region for some confidence level.

Proof of Theorem 5.1. Recall that $\Pi_{\lambda,a} = \{\pi | \pi$ is the indefinite integral over (β,σ) of a measurable function, $f(\beta,\sigma)$, such that $\sigma^{-(K+2)} \leq f(\beta,\sigma) \leq (1+a)\sigma^{-(K+2)}$, for all $(\beta,\sigma)\}$. If $\pi \in \Pi_{\lambda,a}$, let $f_\pi(\beta.,\sigma)$ be the density of π. If $\pi \in \Pi_{\lambda,a}$, then the corresponding posterior probability distribution of β has a density proportional to the integrable function $p_\pi(\beta) = \int_0^\infty \sigma^{-T}\exp\{-(1/2\sigma^2)[(T-K-1)s^2 + (\beta-b)'X'X(\beta-b)]\}f_\pi(\beta,\sigma)d\sigma$. Clearly, if π and π' in $\Pi_{\lambda,a}$ are such that $f_\pi(\beta,\sigma) \leq f_{\pi'}(\beta,\sigma)$, for all β and σ, then $p_\pi(\beta) \leq p_{\pi'}(\beta)$, for all β. Therefore, the set of posterior probability distributions for β corresponding to the set of priors $\Pi_{\lambda,a}$ is $P_{\lambda,a}(\beta) = \{\nu(R^{K+1})^{-1}\nu | \nu$ is the indefinite integral of $p(\beta)$, where $p_\lambda(\beta) \leq p(\beta) \leq (1+a)p_\lambda(\beta)$, for all $\beta\}$. It may be calculated that $p_\lambda(\beta)$ is proportional to $[T-K-1 + [(\beta-b)'X'X(\beta-b)]/s^2]^{-(T+K+1)/2}$.[2]

I now introduce some new notation. If $q : R^{K+1} \to [0,\infty)$ is integrable, let $I_q = \{(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} Q(\beta_0, \ldots, \beta_K)d\beta_0 \cdots d\beta_K)^{-1}$ $(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \beta_0 Q(\beta_0, \ldots, \beta_K)d\beta_0 \cdots d\beta_K, \ldots,$ $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \beta_K Q(\beta_0, \ldots, \beta_K)d\beta_0 \cdots d\beta_K) | Q : R^{K+1} \to [0,\infty)$ is measurable and $q(\beta_0, \ldots, \beta_K) \leq Q(\beta_0, \ldots, \beta_K) \leq (1+a)q(\beta_0, \ldots, \beta_K)$, for all $(\beta_0, \ldots, \beta_K)\}$. Thus, $M_{\lambda,a}(\beta|X,y) = I_{p_\lambda}$.

It is not hard to verify that I_q is convex, for any q. An obvious weak compactness argument implies that I_q is compact.[3]

Now consider a change of variables $\rho = c + C\beta$, where C is a nonsingular $(K+1) \times (K+1)$ matrix. Then, the density $p(\rho)$ corresponding to

---

[2] See Zellner (1971), p. 67.

[3] By the weak topology, I mean the weak topology defined from the duality between the vector space of bounded continuous functions on R^{K+1} and the vector space of signed measures on R^{K+1}.

the density $q(\beta)$ is $p(\rho) = (\det C)^{-1} q(C^{-1}(\rho - c))$. Clearly,

$$(5.2) \qquad I_p = c + CI_q = \{c + Cz \mid z \in I_q\} .$$

is the set of means of ρ .

I now show that if the density q is of the form $q(\beta) = f(\beta \cdot \beta)$, for some function f , then

$$(5.3) \qquad I_q = \{\beta \mid \beta \cdot \beta \leq \gamma^2\} , \quad \text{for some } \gamma \geq 0 .$$

In order to prove that (5.3) is true, let C be any $(K+1) \times (K+1)$ orthogonal matrix and let $\rho = C\beta$. If p is the density of ρ corresponding to q , then $p(\rho) = (\det C)^{-1} q(C^{-1}\rho) = f(C^{-1}\rho \cdot C^{-1}\rho) = f(\rho \cdot \rho)$ $= q(\rho)$, so that by (5.2) $I_q = I_p = CI_q$. Since C is an arbitrary orthogonal matrix and I_q is convex and compact, it follows that I_q is a closed ball about the origin of R^{K+1} . This completes the proof of (5.3).

Since the matrix $B = s^{-2}X'X$ is symmetric and positive definite, there is an orthogonal matrix C such that $C'BC = D$, where D is diagonal. Let $D^{1/2}$ be the unique diagonal matrix such that $D^{1/2}D^{1/2} = D$. Finally, let $\rho = D^{1/2}C'(\beta - b)$. The density $p_\lambda(\beta)$ is of the form $p_\lambda(\beta) = f((\beta-b)'B(\beta-b))$. Therefore, the density, q , of ρ corresponding to p_λ is proportional to $f(\rho \cdot \rho)$. By (5.3), $I_q = \{\rho \mid \rho \cdot \rho \leq \gamma^2\}$, for some $\gamma \geq 0$. It follows that $I_{p_\lambda} = \{\beta \mid (\beta-b)'B(\beta-b) \leq \gamma^2\}$.

In order to evaluate γ , I use a result of DeRobertis and Hartigan (1981) to evaluate $\gamma = \sup\{\rho_0 \mid \rho \in I_q\}$. Clearly, $\gamma = \sup\{(\int_{-\infty}^{\infty} Q(\rho_0)d\rho_0)^{-1} \int_{-\infty}^{\infty} \rho_0 Q(\rho_0)d\rho_0 \mid Q : R \to R$ is measurable and $q_0(\rho_0) \leq Q(\rho_0) \leq (1+a)q_0(\rho_0)$, for all $\rho_0\}$, where

$q_0(\rho_0) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q(\rho_0, \ldots, \rho_K) d\rho_1, \ldots, d\rho_K$. By Theorem 4.1 of DeRobertis and Hartigan (1981), γ is the unique solution of the equation

$$(5.4) \qquad (1+a) \int_{-\infty}^{\infty} (\rho_0 - \gamma)^+ q_0(\rho_0) d\rho_0 + \int_{-\infty}^{\infty} (\rho_0 - \gamma)^- q_0(\rho_0) d\rho_0 = 0 ,$$

where $x^+ = \max(0, x)$ and $x^- = \min(0,x)$.

Using the definition of ρ , one sees that $q(\rho_0, \ldots, \rho_K)$ is proportional to $[T-K-1 + \rho_0^2 + \ldots + \rho_K^2]^{-(T+K+1)/2}$. Integrating with respect to $\rho_1, \ldots, \rho_K$, one obtains that $q_0(\rho_0)$ is proportional to $[T + T\rho_0^2/(T-K-1)]^{-(T+1)/2}$, which is proportional to the density of the Student t distribution with T degrees of freedom for the variable $\sqrt{T/(T-K-1)}\ \rho_0$. The formula for γ in Theorem 5.1 is obtained by substituting this Student t density for $q_0(\rho_0)$ in equation (5.4). Q.E.D.

One can use the Knightian confidence regions to formulate tests of linear hypotheses on the regression coefficients of the normal linear regression model. These tests are of the some form as the classical tests.

Consider the general linear hypothesis $\beta \in V + \bar{\beta}$, where V is a linear subspace of R^{K+1} . The Knightian test of the hypothesis $\beta \in V + \bar{\beta}$ is to accept if $0 \in \pi_{V^c}(M_{\lambda,a}(\beta|X,y) - \bar{\beta})$, where $\pi_{V^c} : R^{K+1} \to V^c$ is the orthogonal projection of R^{K+1} onto the orthogonal complement, V^c , of V in R^{K+1} . For instance, the Knightian test of the hypothesis $\beta_{H+1} = \ldots = \beta_K = 0$ is to accept if $0 \in \pi_2(M_{\lambda,a}(\beta|X,y))$, where $\pi_2 : R^{K+1} \to R^{K-H}$ is the orthogonal projection which carries $(\beta_0, \ldots, \beta_K)$ to $(\beta_{H+1}, \ldots, \beta_K)$. The next proposition makes it possible to compare this test with the classical one. In the proposition, X_1 denotes the matrix consisting of the first H+1 columns of X and X_2 denotes the

last K-H columns. Also, $\beta_{(1)} = (\beta_0, \ldots, \beta_H)'$, $\beta_{(2)} = (\beta_{H+1}, \ldots, \beta_k)$ and $b_{(2)} = (b_{H+1}, \ldots, b_K)'$, where $b = (X'X)^{-1}X'y$.

Proposition 5.5. $\pi_2(M_{\lambda,a}(\beta|X,y)) = \{\beta_{(2)} | (\beta_{(2)} - b_{(2)})'D(\beta_{(2)} - b_{(2)}) \leq \gamma^2 s^2\}$, where $D = X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2$ and where γ is as in Theorem 5.1.

Proof. Clearly, $\pi_2(M_{\lambda,a}(\beta|X,y)) = \{\beta_{(2)} | \text{for some } \beta_{(1)}, (\beta_{(1)}', \beta_{(2)}' - b_{(2)}')X'X\begin{pmatrix}\beta_{(1)}\\ \beta_{(2)} - b_{(2)}\end{pmatrix} \leq \gamma^2 s^2\}$. Using the symmetry of $X'X$, it is not hard to see that $(\beta_{(1)}', \beta_{(2)}' - b_{(2)}')X'X\begin{pmatrix}\beta_{(1)}\\ \beta_{(2)} - b_{(2)}\end{pmatrix}$ $= (\beta_{(2)} - b_{(2)})'D(\beta_{(2)} - b_{(2)}) + (\beta_{(1)} + B_{11}^{-1}B_{12}(\beta_{(2)} - b_{(2)}))'B_{11}(\beta_{(1)} + B_{11}^{-1}B_{12}(\beta_{(2)} - b_{(2)}))$, where $B_{ij} = X_i'X_j$, for $i, j = 1, 2$. Since B_{11} is positive definite, $(\beta_{(2)} - b_{(2)})'D(\beta_{(2)} - b_{(2)}) \leq \gamma^2 s^2$ whenever $(\beta_{(1)}, \beta_{(2)} - b_{(2)})'X'X\begin{pmatrix}\beta_{(1)}\\ \beta_{(2)} - b_{(2)}\end{pmatrix} \leq \gamma^2 s^2$. If $(\beta_{(2)} - b_{(2)})'D(\beta_{(2)} - b_{(2)}) \leq \gamma^2 s^2$, then $(\beta_{(1)}, \beta_{(2)} - b_{(2)})'X'X\begin{pmatrix}\beta_{(1)}\\ \beta_{(2)} - b_{(2)}\end{pmatrix} \leq \gamma^2 s^2$, where $\beta_{(1)} = -B_{11}^{-1}B_{12}(\beta_{(2)} - b_{(2)})$. Q.E.D.

From this proposition, one see that the Knightian test of the hypothesis $\beta_{(2)} = \bar{\beta}_{(2)}$ is to accept the hypothesis if $(b_{(2)} - \bar{\beta}_{(2)})'D(b_{(2)} - \bar{\beta}_{(2)}) \leq \gamma^2 s^2$. The classical test of this test is to accept the hypothesis if $(b_{(2)} - \bar{\beta}_{(2)})'D(b_{(2)} - \bar{\beta}_{(2)}) \leq \delta s^2$, where $F_{T-K-1}^{K-H}(\delta/(K-H)) = 1-\alpha$ and 100α is the significance level of the test and F_{T-K-1}^{K-H} again denotes the cumulative distribution function of the F distribution. Thus, the Knightian and classical tests differ only in the definitions of the numbers γ^2 and δ.

The same relationship exists between classical and Knightian tests in the special case of a point hypothesis about one component of β. Let s_k be the k^{th} diagonal entry of $(X'X)^{-1}$. If $H = K-1$, then $D = s_K^{-1}$. From this fact and proposition 5.5, it follows that the Knightian test of the hypothesis $\beta_k = \bar{\beta}_k$, for some k, is to accept the hypothesis if $(b_k - \bar{\beta}_k)^2 \leq \gamma^2 s^2 s_k$. The classical test is to accept the hypothesis if $(b_k - \bar{\beta}_k)^2 \leq \delta^2 s^2 s_k$, where $F_{T-K-1}(\delta) = 1-\alpha$, F_{T-K-1} being the cumulative distribution function of the Student t distribution with $T-K-1$ degrees of freedom.

It follows easily from proposition 5.5 that the equivalence of classical and Knightian tests applies to tests of linear hypotheses of the form $\beta \in V$, where V is a linear subspace of R^{K+1}. The classical test may be described as follows. Let e be the vector of residuals of the regression of y onto the columns of X and let e_V be the vector of residuals of the regression of y onto the columns of X_V, where the rows of X_V are the orthogonal projections of the rows of X onto V. The classical test is to accept that $\beta \in V$ if $(e_V \cdot e_V - e \cdot e)(T-K-1) \leq \delta(e \cdot e)(K-H)$, where $H+1$ is the dimension of V and $F^{K-H}_{T-K-1}(\delta) = 1-\alpha$. By theorem 5.1, the Knightian test is to accept that $\beta \in V$ if $0 \in \pi_{VC} B_\gamma$, where $B_\gamma = \{\beta \in R^{K+1} \mid (\beta-b)'X'X(\beta-b) \leq \gamma^2 s^2\}$.

Proposition 5.6. For any α such that $0 < \alpha < 1$, there is a $c > 0$ such that $0 \in \pi_{VC} B_c$ if and only if $[(e_V \cdot e_V - e \cdot e)(T-K-1)]/[(e \cdot e)(K-H)] \leq \delta$ where $F^{K-H}_{T-K-1}(\delta) = 1-\alpha$.

Proof. By making an orthogonal change of basis of R^{K+1}, one may assume that $V = \{\beta \in R^{K+1} | \beta_{H+1} = \ldots = \beta_K = 0\}$. Under this hypothesis, it is not hard to see that

$$\frac{(e_V \cdot e_V - e \cdot e)(T-K-1)}{(e \cdot e)(K-H)} = \frac{(\beta_{(2)} - b_{(2)})' D (\beta_{(2)} - b_{(2)})}{s^2 (K-H)},$$

where the notation is as in proposition 5.5. By that proposition, there is $c > 0$ such that

$$\frac{(\beta_{(2)} - b_{(2)})' D (\beta_{(2)} - b_{(2)})}{s^2 (K-H)} \leq \delta$$

if and only if $0 \in \pi_2 B_c$. Q.E.D.

The equivalence between classical and Knightian tests or confidence regions offers one means of reconciling econometric practice with the multiplicity of subjective distributions assumed in Knightian decision theory. The classical confidence region may be interpreted as indicating the size of the set of posterior distributions which should be attached to regression parameters. Significance levels are measures of the level of uncertainty aversion associated with the tests.

The equivalence between classical and Knightian methods also suggests that perhaps practicing econometricians act unconsciously as Knightian decision makers. In support of this suggestion, one can argue that the attitude commonly adopted toward hypothesis testing seems Knightian in spirit. This attitude seems to be that it is reasonable to believe that a hypothesis is true, once it is found acceptable according to a good statistical test. The Knightian theory states a sense in which such belief is reasonable.

The suggestion that practicing econometricians are unconsciously Knightian makes it natural to ask what level of uncertainty aversion correspond to the typical confidence level of 95% in the case of linear regression. According to theorem 5.1, the classical and Knightian confidence regions are of the same form. Using the equation of theorem 5.1, one can determine what level of uncertainty aversion, a, would give the classical region of confidence level 95%. The following table gives values of a so determined at specified levels of K and $T-K-1$.

T-K-1 \ K	1	2	3
10	1,355	6,161	20,406
20	1,206	4,915	7,835

These values are enormous. Their largeness is the inverse of the robustness of posterior means to priors observed by DeRobertis and Hartigan (1981).

In order to grasp the meaning of these numbers, recall that a is defined as $(0.5 + \varepsilon)(0.5) - \varepsilon)^{-1} - 1$, where the interval $[0.5 - \varepsilon, 0.5 + \varepsilon]$ is the set of subjective probabilities attached to an event about whose likelihoods of occurrence and non-occurrence one is equally and totally ignorant. If $a = 1{,}000$, then $\varepsilon = 0.499002$, so that at the levels of a in the table, nearly all probabilities between zero and one are applied to the event.

The large levels of uncertainty aversion implied by the 95% confidence level is evidence that uncertainty aversion may be an important aspect of reality.

6. The Maximum Likelihood Point of View

In interpreting confidence regions, I have assumed that a researcher's ultimate objective is to maximize some payoff function which is linear in future values of y_t. That is, researchers were viewed as providing information potentially useful for practical applications. But researchers seem to be preoccupied with learning the truth as well as with possible applications. Talk of truth suggests Jeffreys' (1961) epistemological derivation of Bayesian statistical methods. However, talk of truth also suggests a decision problem having to do with the allocation of scientific resources. Since such resources are scarce, it makes sense to allocate them to the verification of those hypotheses which seem most likely. Since the subject of economics "wins" only if a hypothesis is verified, it makes sense to bet on the most likely candidate hypotheses. At least at a preliminary stage of research, one should sort out the most promising hypotheses whose relative explanatory power would be compared in later work. This search for prominence I call the maximum likelihood point of view, where by likelihood I mean posterior probability. This point of view leads to another Knightian interpretation of confidence regions and also provides some justification for the practice of choosing the regression with the highest R^2.

Suppose that a long list of K regressors is under consideration and that it is believed that an acceptable model has no more than n of them. Also, suppose that *a priori* one is completely ignorant of the regression coefficients. The problem is to choose the most likely set or sets of regressors.

Let H be the set of subsets of $(1, \ldots, K)$ of size n. Each H in H corresponds to the hypothesis that the set of regressors is those

indexed by the elements of H. Let X_H be the $T \times n$ data matrix of the regressors in H and assume that X_H is of rank n, for all H. y denotes the T vector of observations of the dependent variable and S_H^2 denotes the sum of squared error of the regression of y on X_H, decided by $T-n$. Choosing the regression with the highest R^2 corresponds to choosing the one with lowest S_H^2. Let λ_H be the prior with density $a_H\sigma^{-(n+1)}$ defined over the parameters (β_H, σ_H) of the regression y of X_H. The constant a_H must be chosen so that the prior density of y does not depend on the choice of units of the regressors. This restriction implies that $a_H = A_H|X_H'X_H|^{1/2}$, where A_H is a constant. If all hypotheses are treated symmetrically, the constants A_H should be independent of H and so may be set equal to one. Let λ be the global prior generated by the λ_H and defined on the parameters (β_H, σ_H), for all H. Given λ_H, the distribution of y is As_H^{-T}, where A is a constant independent of H. Thus, given λ, the posterior probability of H is proportional to s_H^{-T}, so that the most likely hypothesis is the one with the lowest s_H^2 and highest R^2. If one is Knightian with priors $\Pi_{\lambda,a}$, then hypothesis H dominates H' if and only if $s_H^{-T} > (a+1)s_{H'}^{-T}$ or $s_H^2 < (a+1)^{-2/T}s_{H'}^2$. This criterion can be viewed as a Knightian interpretation of the usual F-tests used to compare H and H'.[4]

The choice of a model, or a small set of models, with the highest R^2 could be misleading for practical decision making. Since different models could have very different practical implications, good decision making might require consideration of all models.

---

[4]This approach to model selection cannot be carried very far. It gives no guidance as to the choice of n. It is not clear how the constants A_H should depend on n.

Empirical economists seem to prefer classical to Bayesian statistical methods, even though economic theorists tend to assume that all rational people are Bayesian. One can speculate that the preference for classical methods stems from their tendency to indicate which hypotheses are prominent. An honest use of Bayesian methods would probably attach small positive prior probabilities to many hypotheses and make the posterior probabilities of each look discouraging low. From the maximum likelihood point of view, one should not necessarily be disheartened by low posterior probabilities.

Turning now to the interpretation of confidence regions, imagine that a researcher wishes to estimate a vector of parameters, β. In order to indicate the most prominent value of β, a Bayesian would probably report the mode of the posterior distribution of β. The mode could be viewed as the best working hypothesis as to the value of β. A Knightian would have a set of modes to report, one for each of his priors. These choices form a confidence region.

Consider now the normal linear regression model of the previous section. Using the standardized set of priors defined there, one obtains that the set of modes is a classical confidence region. The set of posterior densities on the regression coefficients, β, is the set of densities $p(\beta)$, which, before normalization, satisfy $p_\lambda(\beta) \leq p(\beta) \leq (1+a)p_\lambda(\beta)$, for all β, where $p_\lambda(\beta)$ is proportional to $[(T-K-1)s^2 + (\beta-b)'X'X(\beta-b)]^{-(T+K+1)/2}$. A value β can be a mode for some β if and only if $[(T-K-1)s^2 + (\beta-b)'X'X(\beta-b)]^{-(T+K+1)/2} \geq (1+a)[(T-K-1)s^2]^{-(T+K+1)/2}$. It follows that the set of modes is $\{\beta \mid (\beta-b)'X'X(\beta-b) \leq [(1+a)^{2/(T+K+1)}-1](T-K-1)s^2\}$, which is of the same form as a classical confidence region.

One can again ask which levels of uncertainty aversion, a, give a region identical with the classical 95% confidence region. The table below gives such values. They are very large, just as they were in the corresponding table of Section 5.

T-K-1 \ K	1	2	3
10	66	396	2,563
20	35	142	572

7. Uncertainty about the Likelihood Function

In the previous two sections, it was assumed that the random variables y_t were normally distributed. As this seems a somewhat dubious assumption in the context of economic data, it is interesting to note that the results of the previous two sections may be interpreted as expressing uncertainty about the likelihood function as well as about the value of the parameters.

In order to make this interpretation, let Ω be the state space consisting of all vectors $(\beta,\sigma,y) = (\beta_0, \ldots, \beta_K, \sigma, y_1, \ldots, y_T)$ and assume given the matrix, X, of observations of independent variables. Let Λ be the measure on Ω defined from the diffuse prior λ on (β,σ), with density $\sigma^{-(K+2)}$, and assuming that for each (β,σ), the y_t are independently and normally distributed with means $X_t\beta$ and variance σ^2. That is, Λ has density, $p_\Lambda(\beta,\sigma,y)$, proportional to

$$\sigma^{-(T+K+2)} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - X_t\beta)^2\right] .$$

Taking Λ as a central measure, one can form the set of measures expressing equal ignorance relative to Λ and uncertainty aversion of degree a. This set is $\Pi_{\Lambda,a} = \{\nu | \nu$ is a measure on Ω which is the indefinite integral of a measurable function $f : \Omega \to (0,\infty)$ such that for almost every $(\beta,\sigma,y) \in \Omega$, $p_\Lambda(\beta,\sigma,y) \leq f(\beta,\sigma,y) \leq (1+a)p_\Lambda(\beta,\sigma,y)\}$. Clearly, $\Pi_{\Lambda,a}$ expresses uncertainty about the likelihood function, though the scope of the uncertainty is restricted to a neighborhood of the normal likelihood. The distributions in $\Pi_{\Lambda,a}$ are improper, but the distributions conditional on y are proper and are exactly the posterior distributions obtained in Section 5 from $\Pi_{\lambda,a}$ using the normal likelihood function. Thus, $M_{\lambda,a}(\beta|X,y)$ is the set of conditional means of β obtained from the conditional distributions drawn from $\Pi_{\Lambda,a}$. The results of the previous two sections could have been derived using $\Pi_{\Lambda,a}$ rather than $\Pi_{\lambda,a}$.

8. Discoverability

Until now, I have tried to show that the confidence regions and significance levels of classical statistics may be interpreted as manifestations and measures of Knightian uncertainty. However, these arguments do not explain how Knightian uncertainty could be more than a temporary phenomenon. Numerous theorems demonstrate that the distribution of a stochastic process is learned asymptotically as the number of observations goes to infinite. Since economic life is always generating new observations, will we not eventually know nearly perfectly all random processes governing economic life? One obvious answer is that the probability laws governing economic life are always changing. But then, why are not those changes themselves governed by a stochastic law one could eventually discern?

The answer to these questions seems to be that not every stochastic law can be learned from the data it generates. Of course, any sequence is trivially generated by the stochastic law which attaches probability one to the given sequence, and this law is in some sense learned as one observes the sequence. But learning in this sense would not help one predict future values of the sequence. What one has in mind when speaking of learning a stochastic law is learning regularities useful for prediction. Knightian uncertainty would tend to disappear as data accumulated only if such regularities existed and could be inferred from the data. These properties of a law I refer to as discoverability.

Lack of discoverability leaves one at a loss as to how to proceed. Without it, how is one to determine what subjective distributions to attach to future realizations and how is one to test for discoverability? But whatever the inconvenience, I believe that non-discoverability should be admitted as a possibility. There seems to be no sound reason for believing that economic time series necessarily have discoverable laws, the popularity of time series methods notwithstanding.

The concept of discoverability being vague, I here attempt a rigorous definition. Using the definition, I prove that discoverable laws indeed can be discovered and that there exist non-discoverable ones. The definition may be incomplete in that one can think of laws which are discoverable in the loose sense and yet are not included in the definition. However, I believe the definition includes all laws likely to occur in economics.

The definition of discoverability is broken down into two definitions, which I term strong and weak. The strong one includes stationary processes and those generated by difference equations of finite order and with essen-

tially bounded solutions. Their laws can be discovered from the frequency of occurrence of events. The strong definition includes laws which can be discovered from frequencies only after a transformation of the data, such as first differencing.

Discoverability is defined as applying to the law for a given sequence of numbers. The question is whether a law of evolution can be derived for the sequence. The numbers of the sequence are assumed to be integers or finite decimal expansions of numbers with some bound on the number of significant digits to the right of the decimal point.[5] I will mention later how the definition might be extended to sequences of not necessarily integral numbers.

Let $x = (x_1, x_2, \ldots)$ denote a sequence of integers. One might wish to allow each x_n to have many components, but such a sequence can be represented as a sequence of numbers by writing all the components in succession.

Strong Discoverability

Strong discoverability applies to essentially bounded sequences. A sequence x is defined to be essentially bounded if

$$\lim_{B\to\infty} \liminf_{N\to\infty} N^{-1}\left|\{n \mid 1 \le n \le N,\ |x_N| \le B\}\right| = 1 ,$$

where $|\cdot|$ denotes cardinality as well as absolute value.

The set of integers is denoted Z. For positive integers N and M and for $c \in Z^M$ and $y \in Z$, let

---

[5] In fact, decimal expansions should be expressed as pairs of integers, one before the decimal point and one after.

$$p^M_{x,N}(y|c) = \frac{|(n|M < n \le N,\ (x_{n-M}, \ldots, x_{n-1}) = c,\ x_n = y)|}{|(n|M < n \le N,\ (x_{n-M}, \ldots, x_{n-1}) = c)|}$$

if $|(n|M < n \le N,\ (x_{n-M}, \ldots, x_{n-1}) = c)| > 0$. Otherwise, let $p^M_{x,N}(y|c) = 0$.

Definition. A sequence x has a strongly discoverable law if it is essentially bounded and for sufficiently large M , $\lim_{N\to\infty} p^M_{x,N}(y|c)$ exists, for all $y \in Z$ and $c \in Z^M$.

Example. I now give an example of a sequence with a law which is, in a sense, nowhere discoverable. I define a sequence x of 0's and 1's such that for all M and all $c \in (0,1)^M$, $\limsup_{N\to\infty} p^M_{x,N}(1|c) = 1$ and $\liminf_{N\to\infty} p^M_{x,N}(1|c) = 0$.

Let $p_1, p_2, \ldots$ be any sequence of numbers such that $0 < p_n < 1$, for all n , and $\limsup_{N\to\infty} p_n = 1$ and $\liminf_{N\to\infty} p_n = 0$. Define positive integers $N_1, N_2, \ldots, N_k, \ldots$ and the $x_1, x_2, \ldots, x_{N_k}$ by induction on k as follows. Let $x_1, \ldots, x_{N_1}$ be any finite sequence of 0's and 1's such that for $x = (x_1, \ldots, x_{N_1})$, $|p^M_{x,N_1}(1|c) - p_1| < 1/2$, for all $c \in (0,1)$. This is trivially possible. Having defined N_k and $x_1, \ldots, x_{N_k}$, let $x_{N_k+1}, \ldots, x_{N_{k+1}}$ be any finite sequence of 0's and 1's such that for $x = (x_1, \ldots, x_{N_{k+1}})$, $|p^{K+1}_{x,N_{k+1}}(1|c) - p_{k+1}| < 2^{-k+1}$, for all $c \in (0,1)^{k+1}$. If $x_{N_k+1}, x_{N_k+2}, \ldots$ is chosen according to a binomial process with probability p_{k+1} that $x_n = 1$, then as N_{k+1} goes to infinity, the probability that the sequence $x_1, \ldots, x_{N_{k+1}}$ satisfies the above conditions goes to one. Therefore, there exists an N_{k+1} and a sequence $x_{N_k+1}, \ldots, x_{N_{k+1}}$ satisfying the conditions. This completes the definition of x .

I now show that x has the desired property. Clearly, if $k < K$ and $c \in \{0,1\}^k$, then $|p^k_{x,N_K}(1,c) - p_K| < 2^{-K}$, since $|p^K_{x,N_K}(1,c) - p_K| < 2^{-K}$, for all $c \in \{0,1\}^K$. Therefore, for any k and any $c \in \{0,1\}^k$, $\liminf_{n\to\infty} p^k_{x,n}(1,c) \leq \liminf_{K\to\infty} p^k_{x,N_K}(1,c) = \liminf_{K\to\infty} p_K = 0$ and $\limsup_{n\to\infty} p^k_{x,n}(1,c) \geq \limsup_{K\to\infty} p^k_{x,N_K}(1,c) = \limsup_{K\to\infty} p_K = 1$. This completes the discussion of the example.

I now turn to the sense in which a strongly discoverable law may be discovered. Clearly, no algorithm can determine in finitely many steps whether a sequence is essentially bounded or has a discoverable law, for these are properties of the tail of the sequence. Also, even if it is known that a sequence has a strongly discoverable law, no algorithm can be constructed which will in all cases eventually stop and produce a distribution known to approximate the limit distribution of the sequence to within some preset margin of error. However, one can easily define an algorithm which continues indefinitely and from any sequence with a strongly discoverable law produces from some point or an approximation of the limit distribution which is accurate to within some preset margin of error.

To be more precise, a <u>strong approximation</u> of the law of $x = (x_1, x_2, \ldots)$ is defined to be (M,B,p), where M and B are positive integers and $p : I_B \times I_B^M \to [0,1]$, I_B being $\{-B, -B+1, \ldots, B\}$. The value of p at $(y,c) \in I_B \times I_B^M$ is written as $p(y|c)$. Suppose that x is essentially bounded and has a strongly discoverable law and let $p^M_x(y|c) = \lim_{N\to\infty} p^M_{x,N}(y|c)$, for $(y,c) \in I_B \times I_B^M$. For $\varepsilon > 0$, $\delta > 0$ and positive integer m, the strong approximation (M,B,p) is said to approximate the law of x up to (ε,δ,m) if $M \geq m$, $\limsup_{N\to\infty} N^{-1}|\{n : |x_n| > B\}| < \varepsilon$ and $|p(y|c) - p^M_x(y|c)| < \delta$, for all

$(y,c) \in I_B \times I_B^M$. An approximation algorithm is a calculable function which for each finite sequence of integers $(x_1, \ldots, x_N)$ produces a strong approximation $(M,B,p|(x_1, \ldots, x_N))$.

8.1) Theorem. For any $\varepsilon > 0$, $\delta > 0$ and positive integer m , there exists an approximating algorithm, $(M,B,p)(\cdot)$, such that if $x = (x_1, x_2, \ldots)$ is essentially bounded and has a strongly discoverable law, then for sufficiently large N , $(M,B,p)(x_1, \ldots, x_N)$ approximates the law of x up to (ε,δ,m) and the B component of $(M,B,p)(x_1, \ldots, x_N)$ does not depend on N .

Proof. Suppose the algorithm is presented with $(x_1, \ldots, x_N)$. For $n = 1, \ldots, N$ and $B = 1, \ldots, N$; the algorithm computes $e_{B,n} = n^{-1}|\{k|1 \le k \le n \text{ and } |x_k| \ge B\}$. It then computes, for each $B \in \{1, \ldots, N\}$, $K_{B,N}$, which is the smallest integer K such that $e_{B,n} < \varepsilon$, for $n = K, K+1, \ldots, N$, if such an integer exists. Otherwise, $K_{B,N} = N+1$. Next, the algorithm computes $B_{x,N}$, which is the smallest $B \in \{1, \ldots, N\}$ such that $K_{B,N} \le B$, if such a B exists. Otherwise, $B_{x,N} = N+1$. Finally, the algorithm computes $p^m_{x,N}(y|c)$, for $(y,c) \in I_B \times I_B^m$, and reports $(M,B,p)(x_1, \ldots, x_N) = (m, B_{x,N}, p^m_{x,N})$.

Now suppose that x is essentially bounded and has a strongly discoverable law. Clearly, there is B such that $B_{x,N} = B$, for N sufficiently large. Since $I_B \times I_B^m$ is a finite set and $\lim_{N\to\infty} p^m_{x,N}(y|c) = p^m_x(y|c)$, for all $(y,c) \in I_B \times I_B^m$, $(m, B_{x,N}, p^m_{x,N})$ approximates the distribution of x up to (ε,δ,m) , for N sufficiently large. Q.E.D.

Weak Discoverability

If a sequence diverges to infinity, one cannot expect to derive its law from frequencies alone. Events containing only finitely many points may all occur only finitely often. One might try to derive the law from the frequencies of all subsets of Z^M, for some large M. But this approach is not feasible since there are a continuum of subsets of Z^M. Another strategy is to look at the computable subsets of Z^M, there being only countably many of them. I use a related but less general approach which seems to be closer to statistical practice. The idea is that the law of the sequence $x = (x_1, x_2, \ldots)$ may be discovered if there is a computable sequence of functions, $f(N, x_1, \ldots, x_N)$, for $N = 1, 2, \ldots$, such that the sequence of integers $d_N = f(N, x_1, \ldots, x_N)$ has a strongly discoverable law. The law of x may be recovered from knowledge of f and of the law $d = (d_1, d_2, \ldots)$, provided $f(N, x_1, \ldots, x_N)$ is a one-to-one function of x_N.

The computable functions are, according to Church's thesis, those which are general recursive. (See Monk (1976), Chapter 3 and p. 46.) A difficulty arises immediately because recursive functions are defined to be functions of a fixed number of variables, not of any finite number of variables. Having found no definition for the latter case, I attempt one here.

An integer-valued function $f(N, x_1, \ldots, x_N)$, where the x_N are integers and $N = 1, 2, \ldots$, is said to be <u>recursive</u> if there exist positive integers M and K and general recursive functions, $g_m(N, y_1, \ldots, y_m, x)$, for $m = 1, \ldots, M$, and if there exist initial values $y_{10}, \ldots, y_{M0}$ such that the y_{mN}, for $m = 1, \ldots, M$ and $N = 1, 2, \ldots$ are generated by the inductive formula

$y_{mN} = g_m(N, y_{1,N-1}, \ldots, y_{M,N-1}, x_N)$, for all M and N . Finally, $f(N, x_1, \ldots, x_N) = g_M(N, y_{1,N-1}, \ldots, y_{M,N-1}, x_N)$. Since the functions g_m are general recursive, it is understood that the y_{mN} are integers.

Observe that the above definition of recursiveness is closed under composition. That is, if $f(N, x_1, \ldots, x_N)$ and $g(N, d_1, \ldots, d_N)$ are recursive, then the function $h(N, x_1, \ldots, x_N) = g(N, f(1, x_1), \ldots, f(N, x_1, \ldots, x_N))$ is recursive.

Definition. A sequence $x = (x_1, x_2, \ldots)$ is said to have a weakly discoverable law if there is a recursive function $f(N, x_1, \ldots, x_N)$ such that

1) $d_N = f(N, x_1, \ldots, x_N)$, $N = 1, 2, \ldots$ is an essentially bounded sequence and has a strongly discoverable law and if

2) for all N and $x_1, \ldots, x_{N-1}$, the function $h(x) = f(N, x_1, \ldots, x_{N-1}, x)$ is injective.

Clearly, a sequence with a strongly discoverable law has a weakly discoverable law. Because recursiveness is closed under composition, one does not enlarge the set of discoverable laws if one assumes that the sequence d_N defined above is weakly rather than strongly discoverable.

Remark. The decimal expansions of e and π have weakly discoverable laws.

8.2) Theorem. There is a sequence of integers with no weakly discoverable law.

Proof. Let F be the set of recursive functions $f(N, x_1, \ldots, x_N)$, defined for any integers $x_1, \ldots, x_N$ and for $N = 1, 2, \ldots$. F is a countable set, since the set of general recursive functions is countable, when recursivity is defined in the usual sense. (See Monk (1976), p. 50.) Let $f^{(1)}, f^{(2)}, \ldots$ be an enumeration of all the members of F.

Let $j(t)$ be the largest integer j such that $t \geq j(j+1)/2$ and let $k(t) = t - j(t)(j(t)+1)/2 + 1$. Thus, $k(1) = 1$, $k(2) = 2$, $k(3) = 1$, $k(4) = 2$, $k(5) = 3$, $k(6) = 1$, and so on. The sequence $x = (x_1, x_2, \ldots)$ is defined by induction as follows. Let $x_1 = 1$. Suppose that $x_1, \ldots, x_{2^t}$ have been defined, where t is a non-negative integer. For $n = 1, \ldots, 2^t$, let $d_n = f^{(k(t))}(n, x_1, \ldots, x_n)$. Since the functions $h_n(x) = f^{(k(t))}(n, x_1, \ldots, x_{n-1}, x)$ is injective and maps integers to integers, its range is unbounded. Let x_{2^t+1} be such that $|f^{(k(t))}(2^t+1, x_1, \ldots, x_{2^t+1})| > t$. Having defined $x_{2^t+1}, \ldots, x_{2^t+n}$, for $n < 2^t$, let x_{2^t+n+1} be such that $|f^{(k(t))}(2^t+n+1, x_1, \ldots, x_{2^t+n+1})| > t$. This defines x_N, for $N = 2^t+1, \ldots, 2^{t+1}$.

Suppose that $x_1, x_2, \ldots$ has a weakly discoverable law. Then, for some $f \in F$, $d_N = f(N, x_1, \ldots, x_N)$ is an essentially bounded sequence. But $f = f^{(i)}$, for some i, and by construction for any $B > 0$ $\limsup_{N\to\infty} N^{-1}|\{n : 1 \leq n \leq N \text{ and } |f^{(i)}(n, x_1, \ldots, x_n)| > B\}| \geq 1/2$. Hence, the sequence $d_1, d_2, \ldots$ is not essentially bounded. Q.E.D.

I now consider how a weakly discoverable law may be discovered. A weak approximation of the law of $x = (x_1, x_2, \ldots)$ consists of (f,M,B,p), where $f(N, x_1, \ldots, x_N)$ is recursive in the sense defined above and (M,B,p) is a strong approximation of the sequence $d_N = f(N, x_1, \ldots, x_N)$, for $N = 1, 2, \ldots$. An approximating algorithm, (f,M,B,p), produces for

each finite sequence of integers $(x_1, \ldots, x_N)$ a weak approximation, $(f,M,B,p)_{(x_1,\ldots,x_N)}$, of the law of x.

8.3) Theorem. For $\varepsilon > 0$, $\delta > 0$ and any positive integer m, there exists an approximating algorithm (f,M,B,p) such that if $x = (x_1, x_2, \ldots)$ is any sequence with a weakly discoverable law, then for sufficiently large N,

1) the M component of $(f,M,B,p)_{(x_1,\ldots,x_N)}$ is m,

2) $(f,M,B,p)_{(x_1,\ldots,x_N)}$ does not depend on N and

3) the B and p components of $(f,M,B,p)_{(x_1,\ldots,x_N)}$ satisfy $N^{-1}|(n : 1 \le n \le N \text{ and } |d_n| > B)| < \varepsilon$, $|p^m_{d,N}(y|c) - p(y|c)| < \delta$, for any $(y,c) \in I_B \times I_B^m$, where $d = (d_1, d_2, \ldots)$ is the sequence defined by $d_n = f(n, x_1, \ldots, x_n)$.

Remark. The algorithm does not necessarily find the true f associated with x.

Proof of Theorem. Let F be as in the previous proof and let $G = \{(f,B) | f \in F \text{ and } B = 1, 2, \ldots\}$. G is countable. Let $g^{(1)}, g^{(2)}, \ldots$ be an enumeration of G. Write $g^{(j)}$ as $(f^{(j)}, B(j))$.

When the algorithm is presented with $(x_1, \ldots, x_N)$, it computes $d_n^{(j)} = f^{(j)}(n, x_1, \ldots, x_n)$ and $e_n^{(j)} = n^{-1}|(k : 1 \le k \le n \text{ and } |d_k^{(j)}| > B(j))|$, for $j, n = 1, \ldots, N$. The algorithm also computes $p^m_{d^{(j)},n}(y|c)$ for each $(y,c) \in I_{B(j)} \times I^m_{B(j)}$, where $d^{(j)} = (d_1^{(j)}, \ldots, d_N^{(j)})$.

For each j such that $1 \le j \le N$, let $k(j,N)$ be the largest integer $k < N$ such that $e_n^{(j)} < \varepsilon$, for $n = k, \ldots, N$, and such that for all $(y,c) \in I_{B(j)} \times I^m_{B(j)}$,

$$\max_{k\le n\le N} p^{m}_{d^{(j)},n}(y|c) - \max_{k\le n\le N} p^{m}_{d^{(j)},n}(y|c) < \delta ,$$

if such an integer exists. Otherwise, let $k(j,N) = N$. Let $j(N)$ be the smallest integer j such that $k(j,N) = \min_{1\le i\le N} k(i,N)$. The algorithm reports $(f,M,B,p)_{(x_1,\ldots,x_N)} = (f^{(j(N))}, m, B(j(N)), p)$, where $p = p^{m}_{d,k}$ with $d = d^{(j(N))}$ and $k = k(j(N),N)$.

It is easy to verify that the algorithm has the required properties.

Q.E.D.

I here make a few comments as to how one might extend the previous definitions and results to the case in which the x_n are continuous variables. In defining strong discoverability, the probabilities $p^{M}_{x,N}(y|c)$ would have to be defined for small intervals y and cubes c. In defining algorithms for discovering laws, one must assume that the data, x_n, are given as decimal expansions with a fixed number of significant digits. One can learn the law only up to the error introduced by the imprecision of the decimal expansion. One can reduce the error by choosing the intervals y and cubes c so that their boundaries avoid areas where the data tend to accumulate.

In defining weak discoverability, the function f should be defined only for finite decimal expansions of the data. The value of f should also be some finite decimal expansion. In fact, f should be defined as a sequence of functions, $f_1, f_2, \ldots$, where f_i gives the decimal expansion of f up to i significant digits. The function f could also depend on continuous parameters $\theta_1, \ldots, \theta_K$, which again can be expressed only as finite decimal expansions. In order to be able to learn the law of x, f must depend continuously on the x_n and θ_k. More precisely, if x_{ni} and θ_{ki} are the decimal expansions of x_n and θ_k, respectively, up to i

significant digits, then $f(N, x_1, \ldots, x_N, \theta_1, \ldots, \theta_K)$ $- \lim_{i\to\infty} f_i(N, x_{1i}, \ldots, x_{Ni}, \theta_{1i}, \ldots, \theta_{Ki})$ should exist. Under these conditions, the θ_k could be estimated from the data.

In what follows, I will refer to distributions having no discoverable law. By such a distribution, I mean one whose realizations almost surely have no discoverable law.

9. Tests for Discoverability

One can test the hypothesis that data have no discoverable law only against hypotheses containing some specific structure. In order to carry out such a test in a Bayesian or Knightian way, one needs to know what distribution to assign to the data under the hypothesis that it has no discoverable law. According to the results of the appendix, such a distribution should be in the weak closure of either the convex set or convex cone generated by the set, H_0, of probability distributions with no discoverable law. Unfortunately, the weak closure of H_0 is the set of all probability distributions, since lack of discoverability is a property only of the tail of a sequence. Thus, the appendix is no help at all.

One can argue loosely as follows. Lack of discoverability has to do with instability of structure. The structure could be long-run averages or the parameters of a model. Hence, a distribution associated with non-discoverability should be such that averages and parameter estimates are unstable. If any particular model shows structural instability according to some classical test, one may interpret the test as favoring non-discoverability over the particular model being tested.

Imagine one has a sequence of discoverable structural models in mind,

H_1, H_2, ... , to each of which a positive prior probability is attached. One also attaches positive prior probability to non-discoverability, H_0, which amounts to giving positive prior weight to some subjective distribution over the data, this distribution being associated with H_0.[6] Given data, the posterior probability of each hypothesis may be calculated. Failure of stability by any of the models H_1, H_2, ... might tend to favor the posterior probability of the other models and should also favor H_0. If a great deal of data were available and all the models H_1, H_2, ..., H_N, for N large, showed structural instability, one imagines that H_0 would be greatly favored.[7] Eventually, one might learn that one should stop looking for structure.

If one of the models were a linear regression model, a test for structural stability would be a Chow test for the stability of the regression coefficients or any test for heteroskedasticity of the error terms. Failure to pass these tests should probably be thought of as favoring the possibility of no structure as well as showing the need to search for another structure.

One often hears or reads that economic time series are generated by some vector autoregressive (VAR) process, but that the coefficients of the process change over time. The instability of typical VAR coefficients gives weight to the idea that there may be no useful structure behind the data. To say that coefficients are moving arbitrarily is indistinguishable from

---

[6] Such associated distributions are explained at the end of the appendix.

[7] It needs to be proved that a subjective distribution on the data exists which would asymptotically separate H_0 from H_1, ..., H_N, if H_0 were true. Also, one should supply a proof of the connection between classical stability tests and a Bayesian test for H_0.

saying that there is no discoverable law. It must be possible to fit any data with a VAR if the coefficients themselves are allowed to fluctuate, even if these fluctuations are fairly rare. One must question the predictive value of such fluctuating models. Coefficient changes are often justified by referring to some major economic event, but do not such events occur frequently? If one allows coefficien variability, one has a discoverable structure only if the changes in coefficients themselves obey some discoverable law, as in Kalman filter models. But for such models to have passed a real test, one must have enough data to test the structural stability of the process governing the evolution of the coefficients. Ultimately, one should have finitely many stable parameters. Until one does, one cannot say that Knightian uncertainty has been eliminated.

If one concludes that a body of data does not have a discoverable law, it does not follow that one can say nothing about the distribution generating the data. One could assert that the distribution has a temporary structure which changes slowly, but otherwise unpredictably. Or one could estimate bounds on the distribution. The bounds on the probability of an event would be estimates of the <u>limits infimum</u> and <u>limits supremum</u> of the frequency of the event's occurrence, these limits being possibly distinct. Similarly, bounds on the mean of a sequence $x_1, x_2, \ldots$ would be estimates of $\liminf_{N\to\infty} N^{-1} \sum_{n=1}^{N} x_n$ and $\limsup_{N\to\infty} N^{-1} \sum_{n=1}^{N} x_n$. Such estimates might be $\min_{\log N<K\leq N} K^{-1} \sum_{n=1}^{N} x_n$ and $\max_{\log N<K\leq N} K^{-1} \sum_{n=1}^{N} x_n$. The presence of temporary structure or the existence of bounds on a distribution do not preclude the presence of persistent Knightian uncertainty.

10. The Kalman Filter

It is contrary to common sense to assume that the entire economic system is without structure, there being obvious relationships between many economic variables. But there seems to be no reason to assume that these relationships have a permanent form. A good hypothesis might be that they change slowly and that the changes themselves have no discoverable law. If such were the case, there would always be a certain amount of Knightian uncertainty associated with the relationships, no matter how long they had been observed.

This uncertainty can be expressed to some extent by means of the Kalman filter model. This model generates a probability distribution over a sequence of slowly evolving regression parameters. One can use this distribution as the central distribution in the standardized set of distributions defined in Section 3. This set of distributions expresses uncertainty associated with the unpredictability of the parameters.

To be more specific, let $y_t = X_t\beta_t + \varepsilon_t$, $t = 1, 2, \ldots$, be a sequence of regression models, where each X_t is a $K+1$ vector of exogenous variables, β_t is a $K+1$ vector of parameters, and $\varepsilon_1, \varepsilon_2, \ldots$ is a sequence of independently and normally distributed variables with mean zero and variance σ^2. According to the Kalman filter model, the β_t are governed by a process $\beta_t = \beta_{t-1} + \theta_t$, for $t = 2, 3, \ldots$, where the θ_t are independently and normally distributed $K+1$ vectors with mean zero and variance-covariance matrix $\sigma^2\Phi$. Also, the variables $\theta_1, \theta_2, \ldots$ and $\varepsilon_1, \varepsilon_2, \ldots$ are all mutually independent. If one assumes that β_1 is normally distributed with mean $\bar{\beta}$ and variance-covariance matrix V, then for each $T \geq 1$, the distribution of β_{T+1} conditional on $y_1, \ldots, y_T$

and $x_1, \ldots, x_T$ is normal with mean $b_{T,T+1}$ and variance-covariance matrix $V_{T,T+1}$, where $b_{T,T}$, $b_{T,T+1}$, $V_{T,T}$ and $V_{T,T+1}$ are defined by induction on T as follows. $b_{01} = \bar{\beta}$, and $V_{01} = V$. Given $b_{T-1,T}$ and $V_{T-1,T}$, for $T \geq 1$, then $V_{T,T} = (X_T X_T' + V_{T-1,T}^{-1})^{-1}$ and $b_{T,T} = V_{TT}(V_{T-1,T}^{-1} b_{T-1,T} + y_T X_T)$. Given b_{TT} and V_{TT}, for $T \geq 1$, then $b_{T,T+1} = b_{TT}$ and $V_{T,T+1} = V_{TT} + \Phi$.

Let λ be the probability distribution over $(y_1, \beta_1, y_2, \beta_2, \ldots)$ determined by the above model and assuming knowledge of $X_1, X_2, \ldots$. Treat λ as a central distribution expressing the idea that the β_t evolve slowly. Then, $\Pi_{\lambda,a}$ is a set of distributions expressing Knightian uncertainty about the distribution of $(y_1, \beta_1, y_2, \beta_2, \ldots)$, where $\Pi_{\lambda,a}$ is defined as in Section 3. The corresponding set of conditional distributions of β_{T+1}, given knowledge of $y_1, \ldots, y_T$, is $\Pi_{\lambda(T+1),a}$, where $\lambda(T+1)$ is the conditional distribution of β_{T+1}, given $y_1, \ldots, y_T$, according to the distribution λ. The set means of β_{T+1}, for all the distributions in $\Pi_{\lambda(T+1),a}$, is the confidence region $M_{\lambda,a}(\beta_{T+1} | y_1, \ldots, y_T, X) = \{\beta | (\beta - b_{T,T+1})' V_{T,T+1} (\beta - b_{T,T+1}) \leq \gamma\}$, where γ is the unique solution of the equation $\left(\frac{a+1}{a}\right)\gamma = f(\gamma) + \gamma F(\gamma)$, f being the density and F the cumulative distribution function of the normal distribution with mean zero and variance one.

As T goes to infinity, the size of $M_{\lambda,a}(\beta_{T+1} | y_1, \ldots, y_T, X)$ will not converge to zero, so that uncertainty is permanent. This permanence is due to the fact that one assumes uncertainty about the probability distribution of β_{t+1} given β_t as well as about the prior distribution of β_1, σ^2 and Φ. The uncertainty about the evolution of β_t reflects the hypothesis that the changes in β_t have no discoverable law. If one had

enough data, one could test this hypothesis and, if it were accepted, estimate bounds on the distribution of $\beta_{t+1} - \beta_t$.

Remark. One cannot use an invariant distribution for β_t, σ^2 and Φ, for if one does so the posterior distribution of the β_t is not integrable.

11. Conclusion

A plausible picture of the economic world is as follows. Economic time series are related to each other by structures, which may be changing slowly. Confidence regions can express, in a systematic way, Knightian uncertainty about the parameters of those structures. One could express uncertainty about the structural model itself as intervals of posterior probabilities, the probabilities being those attached to various alternative models. But one cannot express this uncertainty in a standardized way. There is no reason to believe that the slow evolution of the structural models itself obeys a discoverable law. Nor is there any reason to believe that discoverable laws drive the common fluctuations in the time series. In this view, there is ample room for Knightian uncertainty about economic environments, even about one well-studied by econometricians.

APPENDIX

Subjective Distributions and Hypotheses

Much of what has been presented in this paper involves subjective prior distributions over distributional hypotheses. The use of such prior distributions involves an ambiguity. In the framework of Knightian and Bayesian decision theory, subjective distributions are parameters of a preference ordering over bets. One can imagine making real bets only on observable events. One cannot observe which of several possible objective distributions is the correct one. So what is the status of prior distributions? This difficulty has been raised by Marshak. (See Marshak et al. (1975).) The commonly accepted answer to his query seems to be that one forms preferences for bets on hypotheses while pretending that the bets could be carried out. (See Marshak, et al. (1975).) Since I find this answer unsatisfactory, I propose another one here. I show that if an additional behavioral assumption is made, then one can derive the prior distributions over hypotheses directly from the subjective distributions over observable data. The additional assumptions is that one prefers lottery A to B if A has higher expected value than B under each of the distributions believed possible.

Consider, first of all, the simple case in which the data are bounded. Letting Ω denote the set of all possible data, assume that Ω is a compact metric space. Let the set of lotteries be $C(\Omega)$, the set of continuous functions on Ω. Give $C(\Omega)$ the supremum norm, $\|\cdot\|$. That is $\|f\| = \max_{\omega\in\Omega} |f(\omega)|$. The set of continuous linear functionals on $C(\Omega)$ is

$rca(\Omega)$, which is the set of regular, countably additive set functions on the Borel measurable subsets of Ω (see Dunford and Schwarz (1957), p. 262). The weak topology or $rca(\Omega)$ is the weakest such set that a sequence ν_n in $rca(\Omega)$ converges to ν if and only if $\lim_{n\to\infty} \int x(\omega)\nu_n(d\omega) - \int x(\omega)\nu(d\omega)$, for all $x \in C(\Omega)$. The weak topology is itself a metric topology (see Parthasarathy (1967), p. 43). Let $\Delta(\Omega) = \{\nu \in rca(\Omega) \mid \nu(\Omega) = 1$ and $\nu(A) \geq 0$, for all Borel measurable $A\}$. A Knightian preference ordering $\succsim$ on $C(\Omega)$ is one such that there is a non-empty, compact and convex subset, Π , of $\Delta(\Omega)$ such that for all x and y in $C(\Omega)$, $x \succsim y$ if and only if $E_\pi(x-y) > 0$, for all $\pi \in \Pi$.

Suppose it is believed that the true distribution of the data lies in a non-empty, compact subset, H , of $\Delta(\Omega)$. Since H is itself a compact metric space, we may define $C(H)$ and $\Delta(H)$ just as $C(\Omega)$ and $\Delta(\Omega)$ were defined. Define the map $F : \Delta(H) \to \Delta(\Omega)$ as follows. If $x \in C(\Omega)$ and $\alpha \in \Delta(H)$, then

$$F(\alpha)\cdot x = \int_H \int_\Omega x(\omega)\nu(d\omega)\alpha(d\nu) .$$

F is continuous with respect to the weak topologies on $\Delta(H)$ and $\Delta(\Omega)$, for if α_n is a sequence in $\Delta(H)$ converging to α and if $x \in C(\Omega)$, then $\int_\Omega x(\omega)\nu(d\omega)$ depends continuously on ν , so that

$$\lim_{n\to\infty} \int_H \int_\Omega x(\omega)\nu(d\omega)\alpha_n(d\nu) = \int_H \int_\Omega x(\omega)\nu(d\omega)\alpha(d\nu) .$$

Since F is continuous, $\Pi_H = F^{-1}(\Pi)$ is compact. Since F is affine, Π_H is convex. The probabilities in Π_H are subjective distributions over H , and if $\Pi_H \neq \phi$, then Π_H defines a Knightian preference ordering over $C(H)$. If $F(\Pi_H) = \Pi$, then one could say that each subjective distribu-

tion in Π could be obtained from one over H. If ν is a probability distribution over Ω and $x \in C(\Omega)$, let $E_\nu x = \int x(\omega)\nu(d\omega)$.

Representation Assumption. If $E_\nu x > 0$, for all $\nu \in H$, then $x \gtrsim 0$.

A.1) Theorem. $F(\Pi_H) = \Pi$.

Proof. It is sufficient to show that Π is contained in the range of F. Suppose that $\pi \in \Pi$ is not in the range of F. Since F is continuous and defined on a compact set, its range is compact. Since π is not in the range of F, $K = \{t\pi | t \geq 0\}$ does not intersect the range of F. Apply the separation theorem (Dunford and Schwarz (1957), p. 417), to K and the range of F in the space $rca(\Omega)$ with the weak topology and with dual space $C(\Omega)$ (Dunford and Schwarz (1957), p. 421). One thus establishes that there is $x \in C(\Omega)$ and a positive number r such that $E_\pi x \leq 0 < r \leq E_\mu x$, for all μ in the range of F. It follows that $E_\nu x \geq r > 0$, for all $\nu \in H$, so that $x \gtrsim 0$, by the representation assumption. Therefore, $E_\pi x > 0$, by the definition of Π. This contradiction proves the theorem. Q.E.D.

Remark: It is not hard to see that the $F(\Pi_H)$ is the weakly closed convex hull of H.

I now turn to the case in which the data are unbounded. Let the set of possible data, Ω, be a locally compact metric space. Let $L(\Omega)$ be the set of continuous functions on Ω with compact support. That is, $L(\Omega) = \{x : \Omega \to (-\infty,\infty) | x \text{ is continuous and } \{\omega | x(\omega) \neq 0\} \text{ is contained in a compact subset of } \Omega\}$. Give $L(\Omega)$ the supremum norm. With this norm, $L(\Omega)$ is a normed vector space, though certainly not a Banach space. The

set of continuous linear functionals on $L(\Omega)$ is $ca(\Omega)$, the set of countably additive set functions on the Borel σ-ring of Ω (Halmos (1950), p. 247). If A is a bounded, Borel measurable subset of Ω, then $|\mu(A)| < \infty$, for all $\mu \in ca(\Omega)$. Each $\mu \in ca(\Omega)$ acts on $L(\Omega)$ via integration. Similarly, each $x \in L(\Omega)$ acts as a linear functional on $ca(\Omega)$. The weak topology on $ca(\Omega)$ is the weakest such that each $x \in L(\Omega)$ is continuous as a functional on $ca(\Omega)$. Let $ca_+(\Omega)$ denote the set of non-negative set functions in $ca(\Omega)$.

Regarding $L(\Omega)$ as the set of lotteries, a preference ordering $\succsim$ on $L(\Omega)$ is Knightian if there is a weakly closed, convex cone, $\Pi \subset ca_+(\Omega)$ such that $x \succsim y$ if and only if $\int(x(\omega) - y(\omega))\mu(d\omega) > 0$, for all $\mu \in \Pi$.

Even though subjective distributions may be unbounded, think of the data as generated by a probability. Let $\Delta(\Omega)$ denote the set of probabilities defined on the Borel σ-ring of Ω, so that $\Delta(\Omega)$ is a subset of $ca_+(\Omega)$. Suppose it is believed that the true probability law belongs to some weakly closed, non-empty subset H of $\Delta(\Omega)$. Let C_H be the weakly closed, convex cone in $ca_+(\Omega)$ generated by H. The object is to find conditions under which $\Pi \subset C_H$.

Strong Representation Assumption. If $x \in L(\Omega)$ and $\int x(\omega)\mu(d\omega) \geq 0$, for all $\mu \in H$, and $\int x(\omega)\mu(d\omega) > 0$, for some $\mu \in H$, then $x \succsim 0$

A.2) Theorem. $\Pi \subset C_H$.

Proof. If the theorem is false, there is $\pi \in \Pi$ such that $\pi \notin C_H$. Separating π from C_H in the space $ca(\Omega)$ endowed with the weak topology, one establishes that there is $x \in L(\Omega)$ such that $\int x(\omega)\pi(d\omega) < 0 \leq \int x(\omega)\mu(d\omega)$, for all $\mu \in C_H$. Let $y \in L(\Omega)$ be such that $y \geq 0$ and $\int y(\omega)\mu(d\omega) > 0$, for some $\mu \in H$. For $\varepsilon > 0$ sufficiently small, $\int (x(\omega) + \varepsilon y(\omega))\pi(d\omega) < 0$ and $\int (x(\omega) + \varepsilon y(\omega))\mu(d\omega) \geq 0$, for all $\mu \in H$, with strict inequality for some μ. Hence, by the strong representation assumption, $x + \varepsilon y \gtrsim 0$ and so $\int (x(\omega) + \varepsilon y(\omega))\pi(d\omega) > 0$, which is a contradiction. Q.E.D.

If H were itself locally compact, one could prove that every π in Π corresponded to a prior distribution on H, as in Theorem A.1. These prior distributions need not be bounded. The diffuse prior of Section 5 would be an example of such a distribution.

Often distributional hypotheses are themselves composites. Thus, H_0 might be the hypothesis that a sequence of random variables has no discoverable law, and H_1 might be the hypothesis that the random variables are independently and normally distributed with common mean and variance. Suppose that $H_0 \cup H_1$ is the entire set of hypotheses. Then, since $\Pi \subset C_{H_0 \cup H_1}$, each probability distribution $\pi \in \Pi$ may be represented as $\pi = \alpha\mu_0 + (1-\alpha)\mu_1$, where μ_i is a probability distribution in C_{H_i}, for $i = 0, 1$, and $\alpha \in [0,1]$ is the prior probability of hypothesis H_0. The distribution μ_i may be thought of as the subjective distribution associated with the hypothesis H_i. Each μ_i is, of course, defined over the data, not over H_i. Similarly, any improper distribution $\pi \in \Pi$ may be written as $\pi = \mu_1 + \mu_2$, where $\mu_i \in C_{H_i}$, for $i = 0, 1$. The μ_i may be thought of as subjective distributions associated with H_i.

REFERENCES

Bewley, Truman (1986). "Knightian Decision Theory: Part I," Cowles Foundation Discussion Paper No. 807, Yale University.

_____ (1987). "Knightian Decision Theory: Part II: Intertemporal Problems," Cowles Foundation Discussion Paper No. 835, Yale University.

DeRobertis, Lorraine and J. A. Hartigan (1981). "Bayesian Inference Using Intervals of Measures," The Annals of Statistics, 9, No. 2, 235-244.

Dunford, Nelson and Jacob T. Schwarz (1957). Linear Operators, Part I: General Theory. New York: John Wiley and Sons.

Halmos, Paul R. (1950). Measure Theory. Princeton: D. von Nostrand.

Jeffreys, Harold (1961). Theorem of Probability, 3rd ed. Oxford: Clarendon.

Knight, Frank H. (1921). Risk, Uncertainty and Profit. New York: Houghton Mifflin.

Leamer, Edward E. (1987). "Econometric Metaphors," in Truman F. Bewley (ed.), Advances in Econometrics. Fifth World Congress. Cambridge, England: Cambridge University Press.

Marshak, Jacob, et al. (1975). "Personal Probabilities of Probabilities," Theory of Decision, 6, 121-153.

Monk, J. Donald (1976). Mathematical Logic. Berlin: Springer-Verlag.

Parthasarathy, K. R. (1967). Probability Measures on Metric Spaces. New York: Academic Press.

Walley, Peter (1984). "Rationality and Vagueness," unpublished manuscript.

Zellner, Arnold (1971). An Introduction to Bayesian Inference in Econometrics. New York: John Wiley and Sons.