Discussion Paper No. 998

BAYESIAN FORECASTING*

by

Ehud Kalai**

and

Ehud Lehrer**

July 1992

---

*The authors wish to thank Michael Rubinovitch for helpful suggestions.

This research was partly supported by Grant No. SES-9022305 from the National Science Foundation, Economics Program.

**Department of Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management, Northwestern University, 2001 Sheridan Road, Evanston, Illinois 60208.

Abstract

Let $X = (X_1, X_2, \ldots)$ be a sequence of random variables distributed according to an unknown distribution $\mu = \mu_X$, and let $\tilde{\mu} = \tilde{\mu}_X$ be a known hypothetical distribution. The paper provides a condition under which the conditional hypothetical distribution, $\tilde{\mu}_{X_{n+1}|X_1,\ldots,X_n}$, approaches the real one, $\mu_{X_{n+1}|X_1,\ldots,X_n}$, and thus making the Bayesian forecasting asymptotically accurate.

1. Introduction

Let $X = (X_1, X_2, \ldots)$ be a sequence of real valued random variables with $\mu = \mu_X$ denoting their unknown joint distribution. For times $n = 1, 2, \ldots$ we denote the history at time n by $X_n^- = (X_1, X_2, \ldots, X_n)$ and the infinite future by $X_n^+ = (X_n, X_{n+1}, \ldots)$. The problem of forecasting we are interested in, is predicting the probability distribution of the future after observing a sufficiently long past, e.g., approximating the conditional probability distributions $\mu_{X_{n+1}|X_n^-}$ or $\mu_{X_{n+1}^+|X_n^-}$. The method of approximation we study starts with a known hypothetical distribution for X, $\tilde{\mu}$, and uses $\tilde{\mu}_{X_{n+1}|X_n^-}$ and $\tilde{\mu}_{X_{n+1}^+|X_n^-}$ to approximate the unknown correct conditional distributions.

Blackwell and Dubins (1962) showed[1] that if μ is absolutely continuous with respect to $\tilde{\mu}$ then with μ-probability one $\tilde{\mu}_{X_{n+1}^+|X_n^-}$ must "merge" with $\mu_{X_{n+1}^+|X_n^-}$. In other words, posteriors computed according to a incorrect distribution will eventually approximate the correct posteriors provided that absolute continuity holds.

The above result has proven useful for applications in game theory and economics.[2] For example, in Kalai and Lehrer (1990b), X_n denoted the action taken at time n by the opponents of a player in an infinitely repeated game. The Blackwell-Dubins theorem was used to illustrate that such a player learns with time to predict the distribution of his opponents' future actions. As a result, it was shown that utility maximizing players in such a game, who start with individual subjective beliefs about opponents' strategies, must converge with time to a true Nash equilibrium play. A

---

[1] See also Diaconis and Freedman (1986) and Schervish and Seidenfeld (1990).

[2] See, for example, Kalai and Lehrer (1990b) and (1992), Nyarko (1992), and Kreps and Fudenberg (1992).

similar approach was used in Kalai and Lehrer (1992) to illustrate convergence to equilibrium of a dynamic economy, by utility-maximizing price-taking agents that learn to forecast future prices given by a vector X_n^+.

It turns out that for economic applications, as above, learning in the Blacwell-Dubins sense is stronger than necessary. Economic agents, who discount future payoffs, have no need to predict the probability of events in the unbounded infinite horizon. Thus, rather than estimating the conditional distribution of the infinite vector X_{n+1}^+, they need only estimate the conditional probabilities of finite horizon futures of the type $X_{n+1}, X_{n+2}, \ldots, X_{n+\ell}$. Also in the economic applications above, the assumption of absolute continuity is strong and severely restrict the generality of the results. Less restrictive sufficient conditions that result in a weaker notion of learning are the subject of this paper.

2. Examples and Definitions

For $n = 1,2,\ldots$ we let Ω_n be the set of values X_n can take. We assume throughout this paper that it is a finite or countable subset of $\mathbb{R}^m$. We let $\Omega_n^- = \times_{j\le t}\, \Omega_n$ and $\Omega = \times_{n\ge 1}\, \Omega_n$. $\mathcal{F}_n$ denotes the smallest σ-algebra of Ω containing the sets with specified values of X_n. $\mathcal{F}_n^-$ denotes the smallest σ-algebra containing the elements of $\cup_{j\le n}\, \mathcal{F}_j$ and $\mathcal{F}$ denotes the σ-algebra generated by all the cylinder sets of the $\mathcal{F}_n^-$'s.

We let μ and $\tilde{\mu}$ be probability distributions of X. Their interpretations in what follows is that μ is the true one while $\tilde{\mu}$ is an alternative one assumed by an uninformed decision maker.

Definition 2.1: $\tilde{\mu}$ merges with μ if for all $\varepsilon > 0$ with μ-probability one there is a time $n(\varepsilon)$ such that for all $n \geq n(\varepsilon)$

$$(1) \qquad |\mu(X_{n+1}^{\cdot} \in A | X_n^{\cdot}) - \tilde{\mu}(X_{n+1}^{+} \in A | X_n^{-})| < \varepsilon \text{ for every } A \in \mathcal{F}.$$

Remark 2.1: In the definition above, and similar ones that follow, inequality 1 is applied only to the random variables $\mu(X_{n+1}^{+} \in A | X_n^{\cdot})$ when they are well defined, i.e., when $\mu(X_n^{-} = x_n^{-}) > 0$. In these cases it is required that the random variables $\tilde{\mu}(X_{n+1}^{+} \in A | X_n^{-})$ also be well defined ($\tilde{\mu}(X_n^{-} = x_n^{-}) > 0$) and satisfy the inequality.

Definition 2.2: μ is absolutely continuous w.r.t. $\tilde{\mu}$ if for every event $A \in \mathcal{F}$, $\mu(X \in A) > 0 \Rightarrow \tilde{\mu}(X \in A) > 0$.

Theorem 2.1 (Blackwell-Dubins): If μ is absolutely continuous w.r.t. $\tilde{\mu}$ then $\tilde{\mu}$ merges with μ.

Example 2.1: Let each X_n be a Bernoulli random variable assuming the values 0 or 1, and let Ω denote the set of infinite sequences of 0-1's endowed with the usual σ-algebra. For each $\theta \in (0,1)$ let μ_θ be the probability distribution on Ω induced by the sequence of i.i.d. Bernoulli variables with $\mu_\theta(X_n = 1) = \theta$. Any distribution F over the unit interval (0,1), say, with the Borel σ-algebra, induces a distribution for X in the following way: $\mu_F(X \in A) = \int_0^1 \mu_\theta(X \in A) F(d\theta)$.

Let $\mu = \mu_{.5}$ and $\tilde{\mu} = \mu_U$ with U being the uniform distribution on (0,1). μ is not absolutely continuous w.r.t. $\tilde{\mu}$ since the event L, that the long run

average of ones is precisely .5, has $\mu(X \in L) = 1$ and $\tilde{\mu}(X \in L) = 0$. Indeed, merging in the sense of Blackwell-Dubins will fail since with μ-probability one, $\tilde{\mu}(X^{+}_{n+1} \in L | X^{-}_{n}) = 0$ for all n.

As mentioned in the introduction, however, weaker merging (as in Kalai-Lehrer (1990a)), sufficient for economic applications, will occur.

Definition 2.3: $\tilde{\mu}$ merges weakly with μ if for every $\varepsilon > 0$ with μ-probability one there is a time $N(\varepsilon)$ such that for all $n \geq N(\varepsilon)$

$$(2) \qquad |\mu(X_{n+1} \in A | X^{-}_{n}) - \tilde{\mu}(X_{n+1} \in A | X^{-}_{n})| < \varepsilon \text{ for every } A \subset \Omega_{n+1}$$

(see Remark 2.1).

Notice that this definition is equivalent to merging for $n + \ell$ periods for any finite ℓ, i.e., replacing X_{n+1} in (2) by the vector $(X_{n+1},\ldots,X_{n+\ell})$.

In Example 2.1 μ_U merges weakly with $\mu_{.5}$. It seems to suggest that if $\tilde{\mu}$ puts a positive probability on a neighborhood of μ then weak merging occurs. The next example shows that this is not the case, under the following notion of closeness.

Definition 2.4: For every $\varepsilon > 0$ we define a neighborhood of μ as follows:

$$C(\mu,\varepsilon) = \{\mu' : \text{with } \mu\text{-probability 1 for all } n$$
$$|\mu(X_{n+1} = x_{n+1} | X^{-}_{n}) - \mu'(X_{n+1} = x_{n+1} | X^{-}_{n})| < \varepsilon \text{ for all } x_{n+1} \in \Omega_{n+1}\}$$

For example, μ_θ is close to $\mu_{\theta'}$ in Example 2.1 whenever θ is close to

θ'. This is due to the independence of the X_n's there. But, obviously, closeness in the sense of Definition 2.4 is not restricted to i.i.d. random variables.

Example 2.2: Let $(\theta_k)_{k=1,2,\ldots}$ be a sequence of real numbers in the interval $(0,1)$ with $\theta_k \to 1$, and let μ_{θ_k} be as in Example 2.1. In addition, let $1 \leq N_1 < N_2 < N_3 \ldots$ be an increasing sequence of integers, and define a distribution μ_{θ_0} as follows. At each time of the type $n = N_i$, independently of other values of X_j's, $\mu_{\theta_0}(X_n = 1) = \mu(X_n = 0) = .5$. For all other times n, $\mu_{\theta_0}(X_n = 1) = 1$.

To construct a distribution $\tilde{\mu}$ on Ω we start with an infinite sequence $\alpha = (\alpha_k)_{k=0,1,2,\ldots}$ of strictly positive weights summing to 1. We draw an integer $I = 0,1,2,\ldots$ according to the distribution α and then use μ_{θ_I} to randomly select the point in Ω. We let $\tilde{\mu}$ be the probability distribution induced by the above procedure.

Now we let μ be a Dirac measure on Ω assigning probability 1 to the sequence $(1,1,\ldots)$.

Notice that $\tilde{\mu}$ is constructed by assigning strictly positive probabilities to distributions which are arbitrarily close to μ. Yet we have the following.

Claim: There are choices of $N_1, N_2, \ldots$ such that $\tilde{\mu}$ does not merge weakly with μ.

Let 1_n denote the finite sequence consisting of n ones.

For $j = 1,2,\ldots$ let

(3) $$\tilde{P}(I = j|1_n) = \alpha_j \theta^{n}_{n_j} / [\sum_{q=1}^{\infty} \alpha_q \theta^{n}_{n_q} + \alpha_0 (1/2)^{s(n)}]$$

and

(4) $$\tilde{P}(I = 0|1_n) = \alpha_0 (1/2)^{s(n)} / [\sum_{q=1}^{\infty} \alpha_q \theta^{n}_{n_q} + \alpha_0 (1/2)^{s(n)}]$$

with s(n) denoting the number of i's with $N_i \leq n$. Then

$$\tilde{\mu}(X_{n+1} = 1|1_n) = \sum_{j=1}^{\infty} \tilde{p}(I = j|1_n)\theta_{n_j} + \tilde{p}(I = 0|1_n)y(n)$$

where $y(n) = 1/2$ if $n + 1 = N_i$ for some i and $y(n) = 1$ otherwise. To justify our claim it suffices to show that the N_i's can be chosen so that $\tilde{p}(I = 0|1_{N_i - 1})$ are bounded below for all values of i. Considering equality (4) we note that the left side of its denominator approaches zero as $n \to \infty$. Thus, by choosing the N_i's sparsely we can make $\tilde{p}(I = 0|1_{n_i - 1})$ arbitrarily close to 1 for all i.

3. Forecasting Finite Future Events

In this section the distribution of X $\mu = \mu_\theta$, where μ_θ is one chosen from a parameterized family $\{\mu_\theta\}_{\theta \in \Theta}$ according to a probability distribution F defined on a set of parameters Θ. Not knowing the chosen value of θ, a Bayesian forecaster starts with a distribution $\tilde{\mu} = \mu_F$ induced on X by the choice of θ according to F, i.e.,

$$\mu_F(X \in A) = \int \mu_\theta(X \in A) F(d\theta).$$

For every $\varepsilon > 0$ we define as before a neighborhood of θ by

$$C(\theta,\varepsilon) = \{\theta' : |\mu_\theta(X_{n+1} = x_{n+1}|X_n^-) - \mu_{\theta'}(X_{n+1} = x_{n+1}|X_n^-)| < \varepsilon$$
$$\text{for all } n \text{ and all values } x_{n+1} \in \Omega_{n+1}\ \mu_\theta\text{-a.s.}\}.$$

We assume that the topology generated by $\{C(\theta,\varepsilon)\}$ is separable, and we let H be the σ-algebra of Θ, so F is a distribution over (Θ,H).

We assume that: (i) every open set is in H, and (ii) that the functions $g(\theta) = \mu_\theta(A)$ are measureable so μ_F is well defined.

As we know from Example 2.2, there are realizations of θ such that, even if F assigns positive probability to their neighborhood, μ_F will not even weakly merge with μ_θ. However, the set of such θ must have F-measure 0.

Theorem 3.1: Suppose $F(C(\theta,\delta)) > 0$ for all $\theta \in \Theta$ and $\delta > 0$. For F-almost every θ μ_F merges weakly with μ_θ.

Proof: Fix an $\varepsilon > 0$, and fix an open set of the form $C = C(\theta,\varepsilon)$. Denote by ν the measure on $(\Omega,\mathcal{F})$ induced by F restricted to C and by $\{\mu_\theta\}_{\theta\in C}$. That is, $\nu(X \in A) = (1/F(C)) \int_C \mu_\theta(X \in A)dF$, for every $A \in \mathcal{F}$. Since $F(C) > 0$, ν is well defined and, moreover, ν is absolutely continuous w.r.t. μ_F. As a consequence of Blackwell and Dubins' Theorem for $\delta > 0$ with ν probability one there exists a time $N(\delta)$ s.t. if $n \geq N(\delta)$ then

$$|\nu(X_{n+1}^+ \in A|X_n^-) - \mu_F(X_{n+1}^+ \in A|X_n^-)| < \varepsilon/2 \text{ for every } A \in \mathcal{F}.$$

Therefore, there exists a measurable set $C_\varepsilon \subseteq C$ s.t.

(i) $F(C_\varepsilon) = F(C)$, and

(ii) $\forall\ \theta \in C_\varepsilon$ with μ_θ-probability one

$$(5) \qquad |\mu_\theta(X_{n+1} \in A|X_n^-) - \nu(X_{n+1} \in A|X_n^-)| < \varepsilon/2 \text{ for every } A \subset \Omega_{n+1}.$$

Therefore,

$$(6) \qquad |\mu_\theta(X_{n+1} \in A)|X_n^-) - \mu_F(X_{n+1} \in A|X_n^-)| < \varepsilon \text{ for every } A \in \Omega_{n+1}.$$

Since Θ is separable, we conclude that for every $\varepsilon > 0$ and for F-almost every θ it is true that with μ_θ-probability one there is an $N = N(\varepsilon)$ for which (6) holds. By taking a sequence of ε's that go to zero we obtain the theorem.

4. Finite Forecasting with Subjective Assessments

In the previous section it was assumed that the true distribution of X, $\mu = \mu_\theta$, was chosen randomly using a distribution F on the set Θ consisting of all possible values of θ. Theorem 3.1 was motivated by the implicit assumption that F is known to the decision maker so that he can use $\tilde{\mu} = \mu_F$ for his Bayesian updating. In this section we continue with the same model but analyze a decision maker who does not know F. The more general case, where he does not even know that μ is of the form μ_θ, is analyzed first. After that, we consider the special case where he is aware that a distribution of the form μ_θ is correct but does not know the true distribution F by which θ was chosen.

Theorem 4.1: Let $\tilde{\mu}$ be a distribution for X. If $\mu_F << \tilde{\mu}$ and if F assigns positive probabilities to all neighbhorhoods $C(\theta,\delta)$ with $\theta \in \Theta$ and $\delta > 0$, then for F-almost every θ $\tilde{\mu}$ merges weakly with μ_θ.

Proof: By using once more the Blackwell and Dubins' theorem and Theorem 3.1. The details are omitted.

When the decision maker does know that an element of $(\mu_\theta)_{\theta\in\Theta}$ was chosen, but does not know the true distribution F on Θ, he may start with a subjective distribution $\tilde{F}$ on Θ ($\tilde{F}$ is defined on the same σ-algebra H of Section 3). In this case he would use for his Bayesian updating $\tilde{\mu} = \mu_{\tilde{F}}$. It turns out that the absolute continuity assumption can be brought back to F and $\tilde{F}$.

Corollary 4.1: Let $F << \tilde{F}$ and assume that F assigns positive probability to all neighborhoods $C(\theta,\delta)$ for $\theta \in \Theta$ and $\delta > 0$. Then for F-almost every θ $\mu_{\tilde{F}}$ merges weakly with μ_θ.

Proof: It is easy to see under the above assumptions that $\mu_F << \mu_{\tilde{F}}$.

One can restate Corollary 4.1 replacing the absolute continuity assumption by the requirements that F and $\tilde{F}$ are sufficiently defused and that F is "non-atomic."

Corollary 4.2: Suppose $F(D) > 0$ if and only if D contains a neighborhood

$C(\theta,\delta)$ for some $\theta \in \Theta$ and $\delta > 0$ and $\tilde{F}$ assigns positive probability to all such neighborhoods. Then for F-almost every θ $\mu_{\tilde{F}}$ merges weakly with μ_θ.

Proof: Obviously, absolute continuity of F w.r.t. $\tilde{F}$ follows.

References

Blackwell, D. and L. Dubins (1962), "Merging of Opinions with Increasing Information," Annals of Mathematical Statistics, 38, 882-886.

Diaconis, P. and D. Freedman (1986), "On the Consistency of Bayes Estimates," The Annals of Statistics, 14, 1-26.

Kalai, E. and E. Lehrer (1990a), "Weak and Strong Merging of Opinions," to appear in Journal of Mathematical Economics.

Kalai, E. and E. Lehrer (1990b), "Rational Learning Leads to Nash Equilibrium," Northwestern University Discussion Paper.

Kalai, E. and E. Lehrer (1992), "Merging Economic Forecasts," Northwestern University Discussion Paper.

Kreps, D. and D. Fudenberg (1992), "Learning Mixed Equilibria," forthcoming in Games and Economic Behavior.

Nyarko, Y. (1992), "Bayesian Learning Without Common Priors and Convergence to Nash Equilibria," New York University Discussion Paper.

Schervish, M. and T. Seidenfeld (1990), "An Approach to Consensus and Certainty with Increasing Evidence," Journal of Statistical Planning and Inference, 25, 401-415.