Estimating Relative Density on a Metric Space

James MacQueen
University of California, Los Angeles

Abstract. Let $X_1, X_2, \ldots$, be stationary and ergodic random variables with values in a metric space M with distance d, let $P(A) = P(X_n \in A)$ and let $S(x,r) = \{y \in M : d(x,y) \leq r\}$. Let M_0 be the set of x for which $P(S(x,r)) > 0$ if $r > 0$, and suppose also that for x in M_0, $P(S(x,r))$ is continuous in x and is differentiable in r for $r \geq 0$, and with a positive derivative for all r in a neighborhood of 0. Consider the set M^* of pairs (x,y) such that both x and y are in M_0 and $\lim_{r\to 0} P(S(x,r))/P(S(y,r))$ exists and is a finite positive number $R(x,y)$. Then $R(x,y)$ is called the *relative density* of P for the pair x, y.

The differentiability condition is essentially the same as required for P to have a positive density in the Euclidean case. Note there may be pairs of elements (x,y) such that that $\lim_{r\to 0} P(S(x,r))/P(S(y,r)$ fails to exist, is zero, or is $+\infty$. For example, if P on the square $[0,1]X[0,1]$ concentrates a total probability of .5 uniformly on the line $x = y, 0 \leq x, y \leq 1$, and distributes probability uniformly on the square excepting this line, then the line of pairs $x = y$ is in M^* and so is the square excepting the line. But a pair with one element on the line and the other off gives a limit of 0 or $+\infty$ depending of which element appears in the numerator (or denominator). These kinds of measures may be of considerable interest and examples where they arise can be given.

Now let the kernel K be a non-negative, non-increasing real valued function on $[0,\infty)$, and with $\int_0^\infty K(z)dz = 1$. Let $p_{n,b}(x) = \sum_1^n bK(bd(x,X_i))/n$ where $b \geq 0$ is a parameter chosen by the user. For (x,y) in M^*, $R_{n,b}(x,y) = p_{n,b}(x)/p_{n,b}(y)$ is a plausible estimate of $R(x,y)$ and it is shown that as b and n increase without bound, $R_{n,b} \to R$ a.s.

It is intuitive that even when $\lim_{r\to 0} P(S(x,r))/P(S(y,r))$ is 0 or $+\infty$, $R_{n,b}$ will converge a.s. to 0 or $+\infty$ accordingly, but this situation will be treated elswhere.

It is also shown that R provides a workable theory of conditional probability in the general metric space context, without the technical complexities of the Radon-Nykodym approach. A few examples are given of the estimate $R_{n,b}$ illustrating the possible applications including application in psychology

as a model for the development of subjective probabilities from experience.

1. Introduction. This paper develops an extension of the classical kernel estimate of a multivariate density based on i.i.d. variables, to the metric space context using stationary and ergodic variables. An extension of the multivariate model for i.i.d variables to stationary variables has been developed by Rosenblatt(1971).

More specifically, let M, β be a probability space where M is a separable and complete metric space with distance d. Let $X_1, X_2, \ldots$ be stationary and ergodic random variables taking values in M. Let $P(A) = P(X_n \in A)$, $A \in b$, which is independent of n because of the stationarity. Let $S(x,r) = \{y \in M : d(x,y) \leq r\}$ and let M_0 be the set of x in M such that $P(S(x,r)) > 0$ if $r > 0$. Let X be a random variable with distribution P.

The problem is to estimate P using the sample $X_1, X_2, \ldots, X_n$. This formulation includes the problem of estimating the joint probability of X_n, $X_{n+1}, \ldots, X_{n+k}$ for some fixed k, since this sequence can be regarded as single variable Z_n, say, taking values in the product space M^{k+1} and $Z_1, Z_2, \ldots$, itself will be stationary and ergodic. So there is little loss of generality in focusing on the problem of estimating P.

For this problem the only tool that comes readily to mind is the sample frequency function $P_n(A)$, the proportion of the sample X_i in the set A. It is well known that $P_n \to P$ a.s. in the sense of weak convergence: For bounded continuous real valued functions g on M,

$$\sum_1^n g(X_i)/n \to Eg = \int g\,dP$$

a.s. by the ergodic theorem. So $P_n(A)$ is a satisfactory estimate of $P(A)$ for many purposes.

However, there is still much room for improvement of our understanding of this problem. The sample frequency function has little intuitive appeal with very small samples. And perhaps more importantly for this paper, interest in estimating probabilities in a metric space with small samples arises in two other contexts outside of statistics per se. One is in psychology where we ask how do humans acquire subjective probabilities from experience? The other is philosophical: How in some fundamental sense do scientists learn from nature on the basis of observations?

Two important and classical ideas are relevant in these contexts : One is the "continuity of nature" whereby what is learned in one situation can

be transferred to other similar situations, and another is "spread of effect", a term due to Thorndike(1911). In psychology this refers to the fact that if a given signal is followed by a certain stimulus, some expectation is created that if the same or a similar signal occurs again, that same or similar stimulus is likely to follow. Similarly, if a scientist observes a certain sample element, the likelihood of nearby elements is increased which is apparently some sort of spread of effect also. The mind seems to demand this.

The kernel estimate responds well to both these considerations. In fact, both the continuity of nature and the spread of effect are built into the concept of a kernel estimate in a fundamental way.

It might well be asked, "Why study the estimation problem at the level of generality implied by the above metric space formulation?" The answer is that in the social sciences and in psychology in particular, complex objects are studied which cannot be treated by conventional multivariate methods, or if they can be so treated it is only after some data massage which often brings its own problems, such as factor analysis or multi-dimensional scaling. But in nearly all such cases it is possible using human judges to measure similarities and differences in a meaningfull and reproduceble way, and which can be converted to a numerical metric. This is because the human mind is easibly capable of dealing with such objects including comparing them to one another. A musical pattern, or a painting, or a culture, or a war, or a sentence in the English laguage, all seem to the human mind to be recognizable as definite objects, and they can differentiated from one another, rapidly, quickly, and reasonably consistently by trained judges. Their understanding is the essential stuff of these sciences.

The method provided here can be applied in these instances, to in effect, bring them under the perview of an objective statisitical method and even to derrive from such data empirically based predictions just as in conventional time series analysis. The possiblity of measuring distance between such complex objects in a useful and relevant way by a completely objective method is not forclosed.

It is to be noted also, that when such data is collected over time, the standard i.i.d. model will frequently be inapplicable, while in many such instances the stationary model will be reasonably appropriate, although it too will hardly ever be exactly correct.

2. The method. First, the method does not attack the estimation of P directly. Instead the focus is on estimating the relative density R defined in the immediately following.

Let M_0 be the set of x for which $P(S(x,r)) > 0$ if $r > 0$ and suppose also that for x in M_0, $P(S(x,r))$ is continuous in x and is differentiable in r for $r \geq 0$, and with a positive derivative $f(x,r)$ for all r in a neighborhood of 0. Consider the set M^* of pairs (x,y), both x and y in M_0, such that $\lim_{r\to 0} P(S(x,r))/P(S(y,r)$ exists and is a finite positive number $R(x,y)$. Then $R(x,y)$ is called the relative density of P for the pair (x,y). So R is defined only on M^*.

A little consideration shows that the above conditions defining R are just an extension to the general metric space context of the condition on pairs x, y in the Euclidian case, with a density f, such that $f(x)/f(y)$ exists and is finite.

Note there may be pairs of elements x, y such that $\lim_{r\to 0} P(S(x,r))/P(S(y,r)$ fails to exist, is zero, or is $+\propto$. For example, if P on the square $[0,1]X[0,1]$ concentrates a total probability of .5 uniformly on the line of pairs $x = y, 0 \leq x, y \leq 1$, and distributes probability .5 uniformly on the square excepting this line, then the line is in M^* and so is the square excepting the line. But a pair with one element on the line and the other off gives a limit of 0 or $+\propto$ depending of which element appears in the numerator (or denominator). These kinds of measures may be of considerable interest and examples where they arise can be given. The main result here is an estimate $R_{n,b}$ of R based on the sample sequence $X_1, X_2, \ldots, X_n$ and a user chosen parameter b. It is shown (Theorem 1) that if the pair (x,y) is in M^*, $R_{n,b}(x,y) \to R(x,y)$, a.s. as both b and n become large.

Initially, in approaching the problem of estimating a probability distribution in the general metric space context, an attempt was made to define a density for P in the usual way, that is, as a density for P with respect to another convenient measure μ, analogous to Lebesgue measure in Euclidean space, and then to proceed by estimating this density using a kernel method. This attempt failed because it appears that in the general metric space context, there is no natural and easily employed analog of Lebesgue measure. Refocusing the problem on the relative density R, which can be used directly without reference to any other measure than P, leads to a practical theory, including asymptotics, rather easily.

In addition, the relative density provides answers to most questions that the concept of "density" would normally answer and also permits estimates of probabilities and conditional probabilities suitable for many applications. In particular, the relative density offers a simple and useful approach to conditional probabilities though the *conditional odds ratio*, an idea which is explained and defined in Section 5. This notion is quite simple and offers

an alternative to the standard approach to conditional probability using the Radon-Nykodym derivative, which in the problem at hand, does not appear to be operational because, as indicated above, there is no convenient density μ analogous to Lebesgue measure which can be used in the general metric space context.

The estimate $R_{n,b}$ is defined in the following section.

3. The Estimate. Let K be a continuous, positive, non-increasing function on $[0, \infty)$, and with $\int_0^\infty K(r)dr = 1$. Suppose throughout the following the pair (x, y) is in M^* so that $R(x, y)$ is a finite positive number. Let $p(n, b, x) = \sum_{i=1}^n bK(bd(x, Xi))/n$, and let

$$R_{n,b}(x, y) = p(n, b, x)/p(n, b, y). \tag{1}$$

Theorem 1. Let δ, γ be arbitrary positive numbers. Then there is a number b_1 and for each $b \geq b_1$ a number n_1 such that the probability that

$$| R_{n,b}(x, y) - R(x, y) | \leq \delta$$

for all $n \geq n_1$ is at least $1 - \gamma$.

The proof is given in the Appendix.

The conclusion of Theorem 1 continues to hold under the assumption that P concentrates on a countable discrete set D of elements x_i, where $p_i = P(X = x_i) > 0$, $i = 1, 2, \ldots$ and the definition (1) of $R_{n,b}$ is unchanged. The asymptotic behavior of $R_{n,b}(x_i, x_j)$ appears to be almost trivial in this situation. Nevertheless $R_{n,b}$ in the small sample situation appears to be of considerable interest, for x_i and x_j may be near or far, as the case may be, from some number of elements in the training sequence. And surely this should have some effect on the estimated relative likely hood of the x_i and x_j in question. In fact, in the psychological world the spread of effect from a small number of experiences may be of paramount importance. And it appears that such effects are captured by $R_{n,b}$ in a plausible way even with very small samples.

Nevertheless it is desirable to know that the asymptotic condition of b and n increasing without bound still gives the correct result in the limit. Theorem 2 address this question affirmatively, although a mild condition specific to the discrete situation has to be added for the convergence result. This is just that the discrete set D is genuinely discrete, $d(x_i, x_j)$, $x_i \neq x_j$ exceeding some positive number for all such pairs. It could be that the x_i are

all distinct but some sub-sequence of the x_i converge to some x_j, as would be the case, for example, if the elements in M were the countable set of rationals between 0 and 1. The kernel method might work asymptotically in this case but we do not prove any result to this effect. In stead we prove

Theorem 2. Suppose $\inf\{d(x_i, x_j) : x_i \neq x_j\} > 0$. Then if x_i and x_j are in D, and n and b increase without bound, $R_{n,b}\ (x_i, x_j) = p(n, b, x_i)/p(n, b, x_j) \to p_i/p_j = R(x_i, x_j)$ a.s.

The proof is also in the Appendix.

We remark that the parameter b serves the same function as the "window" in the usual kernel estimate of a density and in Euclidean space the method reduces to the usual kernel method as it would be applied to the relative density problem.

To illustrate how the general method operates in a familiar setting, suppose $X_1, X_2, \ldots,$ are i.i.d in E_1 with some unknown density h. Let $b = 1/\sigma$ and let $K(z) = 2\varphi(z), 0 \leq z < \infty$, where φ is the standard normal density. The factor 2 serves to make $\int_0^\infty K(r)dr = 1$. Then, with $d(x, y) = |x - y|$,

$$p(n, b, x) = \sum_1^n bK(bd(x, Xi))/n = \sum_{i=1}^n \left(\frac{2}{\sqrt{\pi\sigma}}\right) \exp(-(1/2)((x - X_i)/s)^2)/n.$$

Except for the factor 2 , this is just the ordinary kernel estimate of the density h at x, using the normal kernel and with "window" σ. So $R_{n,b}(x, y) = p(n, b, x)/p(n, b, y)$ will be an estimate of $h(x)/h(y)$.

In view of the psychological applications b will be called the *spread of effect* parameter in reference to the fact that a response to one stimulus tends to be aroused by other similar stimuli, a central finding from psychology, alluded to above.

While the large sample property of the estimate $R_{n,b}$ given by Theorem 1 is important for the credibility of the method, much of the scientific value of the estimate comes from small sample considerations, and from psychological and philosophical interpretations, although these too are buttressed by the large sample result. This result shows learning by such a method will with sufficient experience lead to good prediction in a very general stationary environment. So the potential domain of applications supported by the consistency result is very large.

4. Interpretation and Application of R. From its definition $R(x, y)$ has the interpretation as the ratio of the probability in a small sphere of radius

r centered at x to the probability of a small sphere, of the same radius, centered at y. This means also that R is an estimate of the relative odds ratio for these two events. That is, given X belongs to the sphere at x *or* the sphere at y, the respective conditional probabilities of these events, say $p(x)$ and $p(y)$, where $p(x) + p(y) = 1$, have the conditional odds ratio $p(x)/p(y)$, which is approximately $R(x, y)$ just from this interpretation. The assumed continuity of $P(S(x, r))$ in r insures that as the radius r of the spheres goes to zero the odds ratio continues to make conceptual sense. In the limit R may be interpreted as the ratio of the probability at the point x to the probability at the point y, which is the usual interpretation of the ratio of the densities at two points, speaking in a certain loose sense, and keeping in mind both events have probability zero. So in this sense R provides a relative density for P at each x, that is, $R(x, y)$ with y fixed may be thought of as an unnormalized density for x. But this usage fails when we attempt to use R to calculate probabilities of larger sets by integration, as though it were truly a density, for at present there is no theory of integration to accompany R.

Nevertheless, the above interpretation of R in terms of ratios of probabilities in small spheres can be extended to estimate discrete conditional distributions and this can be of practical interest. To illustrate, consider selected elements $x_1, x_2, \ldots, x_k$ and some convenient element y. Then

$$p_i = R(x_i, y) / \sum_i^k R(x_i, y) \cong \frac{p(x_i)/p(y)}{\left(\sum_i^k p(x_i)/p(y)\right)} = \frac{p(x_i)}{\sum_i^k p(x_i)}$$

is evidently an estimate of the probability of x_i given some one of the k elements $x_1, x_2, \ldots, x_k$ occurs. If g is a real valued function on M which can be evaluated at each x_i then $\sum_i p_i g(x_i)$ is an approximation to the expected value of g given X falls in the set $\{x_1, x_2, \ldots, x_k\}$.

More importantly, R itself has a *conditional odds ratio* interpretation previously mentioned, and this is defined just below. This can be applied in time series and gives a practical method of calculating from R (or from $R_{n,b}$) estimates of odds ratios for the future given the past and estimates of expected future values of numerical variables.

5. Conditional Odds Ratio. To illustrate, interpret $X_1, X_2, \ldots,$ as successive pairs from a another stationary and ergodic sequence $U_1, U_2, \ldots$ in some metric space M_0 with distance d_0 . That is, let $X_1 = (U_1, U_2)$, $X_2 = (U_2, U_3), \ldots,$ and then each X_i is in $M = M_0 \times M_0$. And in this case it is convenient and useful to take as the metric on M the maximum of the distances in the two component spaces, which is the natural version

of the familiar sup norm for this context. Thus for two elements $x = (r, s)$ and $y = (u, v)$ in M, let $d(x, y) = \max\{d_0(r, u), d_0(s, v)\}$. This metric will be called $\sup d$. Using this distance, $R_{n,b}$ is available as an estimate of R in this context where the distribution of successive pairs is of interest.

Consider then, $R(x, y)$, where $x = (s, u)$ and $y = (s, v)$. Note the first component in each pair is the same element in M_0. This special choice of the pairs is the basis of the conditional odds ratio interpretation, which is now established:

If r is small, then R is approximately the ratio of the probability P of a sphere of radius r around (s, u) to the probability of a sphere of radius r around (s, v) where now P is the distribution of X_n, that is, the pair (U_n, U_{n+1}). Thus

$$\begin{aligned}
R(x, y) &\cong P(X_n \in S(x, r))/P(X_n \in S(y, r)) \\
&= P((U_n, U_{n+1}) \in S(x, r))/P((U_n, U_{n+1}) \in S(y, r)) \\
&= P(d((s, u), (U_n, U_{n+1})) \leq r)/P(d((s, v), (U_n, U_{n+1})) \leq r) \\
&= P(\max\{d_0(s, U_n), \\
&\qquad d_0(u, U_{n+1})\} \leq r)/P(\max\{d_0(s, U_n),\ d_0(v, U_{n+1})\} \leq r) \\
&= P(d_0(s, U_n) \leq r \text{ and } d_0(u, U_{n+1}) \leq r)/P(d_0(s, U_n) \leq r \\
&\qquad \text{and } d_0(v, Un + 1) \leq r).
\end{aligned}$$

Now divide numerator and denominator of this last line by $P(d_0(s, U_n) \leq r)$. Let $S_0(s, r)$ be the sphere of radius r in M_0 centered at s. After the division, the resulting ratio is evidently just the following ratio of conditional probabilities:

$$\begin{aligned}
&R(x, y) \\
&\quad = R((s, u), (s, v)) \\
&\quad \cong P(U_{n+1} \in S_0(u, r) | U_n \in S_0(s, r))/P(U_{n+1} \in S_0(v, r) | U_n \in S_0(s, r)).
\end{aligned}$$

This may be described in words as the conditional odds ratio for the event that U_{n+1} will be in the sphere centered at u relative to the event that U_{n+1} will be in the sphere centered at v, given U_n is in the sphere centered at s.

Such conditional odds ratios permit a variety of applications including rational choice problems where one of several outcomes is in view and they have certain relative utilities and probabilities depending on the action chosen. If the probabilities can be learned from experience, reasonable actions can be selected which maximize, approximately, at least, expected utility.

Moreover, the above interpretation extends immediately to the case

$$X_1 = (U_1, U_2, \ldots, U_{k+1}),\ X_2 = (U_2, U_3, \ldots, U_{k+2}), \ldots,$$

so X_n has values in the product space $M_1 \times M_2 \times \ldots M_{k+1}$ where the spaces may all be different.

Consider then $R((x_1, x_2, \ldots, x_k, x), (x_1, x_2, \ldots, x_k, y))$. Using the metric sup d this becomes the conditional odds ratio for U_{n+k+1} being in a sphere S_{k+1}, in M_{k+1} centered at x, relative to U_{n+k+1} being in a sphere in M_{k+1}, centered at y, both these spheres of radius r, given U_n in S_1, U_{n+1} in $S_2, \ldots, U_{n+k}$ in S_k, where these spheres S_i all have radius r and centers $x_1, x_2, \ldots, x_k$ in the respective spaces $M_1, M_2, \ldots, M_k$.

If the radius r is allowed to go to zero this may be interpreted as the conditional odds ratio for two points on occasion $n + k + 1$ relative to a sequence of past points at times $n, n+1, \ldots, n+k$, these falling in a sequence of different spaces.

This kind of information is of interest because it extends the mathematical language for talking about the regression and prediction problem to the general metric space situation.

While conditional relative odds based on R appear to serve well this purpose, absolute probabilities under P cannot be calculated from R in any obvious way as was previously noted. Nevertheless, there is one condition under which it may be possible to show that R determines P and calculate P from R. This is where the space M and any measurable set in M, can be represented as a union of disjoint spheres of small radii. Then R can be used to find the relative probability of each sphere, and so the normalized sum of the relative probabilities of the spheres whose union is the set in question, can in principle be calculated. But this has not been done in any interesting instance. Moreover, the condition that unions of disjoint spheres provide all the measurable sets and thereby support a general probability measure does not hold in general. Davies(1971) has given an example of a compact metric space with a Borel measure, wherein the measure cannot be captured by its values on disjoint spheres of small radius. So, the relation between R and P remains to be determined.

6. Illustrative applications.

6.1. Choice of the window. Application of the kernel method in requires choice of the window, that is, the parameter b, a problem which is discussed in many papers in the multivariate case. It was difficult to adopt existing methods to the general method under consideration here. So a method had to be devised whose application required only the measure of distance.

One such method is based on "probability concordance", which is to say, agreement, in a specified way, of the kernel estimate of the probabilities of certain sets with their relative frequency in the sample. This method is applicable and has some intuitive appeal even in very small samples. This method is illustrated in the following paragraphs which also illustrates how the method works in practice when applied in the multivariate time series situation.

The method is based on $p(n, b, x)$ and the empirical frequency of the sets S_i of elements within a distance CD of each X_i. There does not appear to be any clear meaning to the raw magnitude of these numbers $p(n, b, x)$, but it seems they ought to be larger in sets of higher frequency.

Accordingly, to evaluate a given b, the average value of $p(n, b, X_i)$ is taken for the sample X_j in S_i, that is, for each set S_i the sum of the $p(n, b, X_j)$ for X_j in S_i is divided by the number of elements in S_i. Then a simple correlation between these averages and the actual number of elements in each set is calculated. This correlation is a raw figure of merit for the value of b in question. Of course, it will also depend on CD. So the procedure is do a direct search over both b and CD and choose the b for which the correlation is a maximum.

This method is used in the following simulation experiment which illustrates conditional odds ratio and the method as well, as applied to a familiar time series problem.

Let $U_1, U_2, \ldots$ be given by $U_{n+1} = .5\ U_n + .1U_{n-1} + \varepsilon_n$ where the disturbances ε_n are i.i.d, and $\varepsilon_n = \pm 1$ with probability $1/2$ each. So the conditional distribution of the next value given the past is sharply bimodal.

A sample sequence of 100 observations ($n = 100$) starting with $U_1 = U_2 = 0$ were generated with this model, and from this training sequence the estimate of R for the space of successive triples $X_n = (U_n, U_{n+1}, U_{n+2})$ was obtained using formula (1) of Section 3 above. The kernel function K was the standard normal $c \exp(-\frac{1}{2} x^2)$ where $c = 2/\sqrt{2\pi}$ is chosen to normalize K, that is, so that $\int_0^\infty K(x)dx = 1$. And the distance was Sup d as applied to the triples.

Using the notation of Section 5 above, we let d_0 be the distance in the space of values of U_n and then for the triples $X_j = (U_j, U_{j+1}, U_{j+2})$, $d(X_j, X_i) = \max(d_0(U_j, U_i), d_0(U_{j+1}, U_{i+1}), d_0(U_{j+2}, U_{i+2}))$.

To choose b, first the elements S_i within a "critical distance" CD of each sample triple (U_i, U_{i+1}, U_{i+2}) were found for $i = 1, 2, \ldots, 98$. So the elements S_i where just those triples $X_j = (U_j, U_{j+1}, U_{j+2})$ in the sample such that $\max(d_0(U_j, U_i), d_0(U_{j+1}, U_{i+1}), d_0(U_{j+2}, U_{i+2})) \leq CD$. The frequency f_i of elements in each of these sets was noted. Then for a given b, the average r_i of $p(n, b, X_j)$ over X_j in S_i was calculated, which is to say r_i is just the sum of such $p(n, b, X_j)$ divided by f_i. Thus from (1) of Section 3,

$$p(n, b, X_j) = (bK(bd(X_j, X_1)) + bK(bd(X_j, X_2)) + \cdots + bK(bd(X_j, X_n)))/n$$

where $d(X_j, X_i)$ is defined just above.

So this gave n pairs of numbers, $f_i,\ r_i$.

It is argued then, that if b is to give a good estimate of the relative density of P, then the f_i and r_i should be correlated. Accordingly the ordinary correlation between the 98 pairs of numbers f_i, r_i was obtained. This is a function of b and CD. Then an "optimal" value of b was found approximately by a direct search over b and CD for the highest correlation. The search was limited to a grid of values of these two parameters. CD ranged from .1 to 1.5 in steps of size .1, and b ranged from 1 to 10 in steps of size 1. A finer grid could be used but this would not serve our illustrative purposes further.

The final choice of b was just the optimal value from this grid search which turned out to be $b = 9.0$.

To illustrate the interpretation and application of this relative density estimate $R_{n,b}$ the conditional odds ratio values for a generic "next" value, called U_3, given a few selected values of the "preceding" pair, called generically U_1, U_2, were calculated, just as described in Section 4, using $b = 9$. This was done for the range of values U_3 from -2.0 to +2.0, in steps of size .2, giving a discrete distribution of the odds ratios, which were then normalized to estimate the distribution of U3 given the selected pair U_1, U_2.

Table 1

Probabilities for Values of Un+1 from Normalized Conditional Odds Ratios

Values	Probabilities
-2.0	.000
-1.8	.003
-1.6	.007

-1.4	.256	
-1.2	.115	
-1.0	.012	
-.8	.000	
-.6	.000	
-.4	.000	
-.2	.000	
0	.000	
.2	.014	
.4	.074	$U_1 = -.727, U_2 = -.289$
.6	.362	
.8	.091	$E_1 = -.302,\ E_2 = -.255,\ E_3 = -.392$
1.0	.001	
1.2	.000	
1.4	.000	
1.6	.000	
1.8	.000	
2.0	.000	

This whole procedure was repeated independently several times with different random seeds for generating the training sequence.

The estimated distribution based on the normalized odds ratios, for a representative run, is shown in Table 1.

It was found that if the given pair U_1, U_2 was not close to any of the training pairs, the predictions could be rather bad, a result which is not surprising. After all, with a purely empirical method, without "structural equations", one should not hope to predict accurately for conditions rarely experienced. For this reason the pair U_1, U_2 actually used in preparing Table 1 was taken as the metric space "center" of the sample training pairs, (U_i, U_{i+1}), $i = 1, 2, \ldots, 99$. The "center" is defined as the pair U_1, U_2 minimizing

$$\sum_{i=1}^{99} \max(d_0(U_1, U_i),\ d_0(U_2, U_{i+1})).$$

So again Sup d is used. This pair is shown in the table.

Notice that the bimodal property of the distribution of the next value of the process U_3 given the preceding two values U_1 and U_2 is very apparent. The predictions can also be evaluated in terms of their mean values. These are given in the table for the following: E_1, the sample average of the scale values $-2, -1.8, \ldots, 1.8, 2.0$ using the probability esti-

mates shown in the table. E_2, the sample average of the sample values $U_1, U_2, U_3, \ldots, U_{100}$. E_3, the theoretical expected value of U_3 given the test input values, $U_1 = -.727$ and $U_2 = -.282$ and the known formula for the process. That is, $E_3 = .5U_1 + .1U_2 = -.392$.

These are roughly what they should be. Other similar calculations gave similar results.

No formal goodness of fit tests were attempted.

It appears also that the method of choosing b worked satisfactorily in this instance.

6.2. A psychological example. This example illustrates the potential application of the kernel estimate $R_{n,b}$ in a simple kind of experimental situation with a more psychological flavor. It is meant to illustrate potential application in psychology. It is also meant to illustrate that it is possible to a analyze very small samples, and that probability concordance may work in this context.

The sample sequence in this instance is five letter non-sense words, each word using letters from the first nine letters of the alphabet. Distance between two words in this space is taken to be the number of letters in one word but not the other, which is also the number of elements in the symmetric difference between two words regarded as five element sets.

To generate the sample words, a fixed five letter word was taken as a prototype. This word was subjected to a random transformation to produce a new word. And this was repeated, independently, starting with the same prototype, five times. Thus the result is five i.i.d five letter words. The whole procedure was repeated twice to produce the two experiments reported in Table 2.

The random transformation was defined by taking a fixed mapping of the set of the first nine letters onto itself. Then two of the elements in the range of this mapping were selected randomly and the values of the map for these two were interchanged , thus making a random map. Then this was repeated, that is , two letters in the range of this map were selected randomly and their values interchanged. Then this map was applied to each of the letters in the prototype word, and the value of the map on these letters became the letters in the new word. Two such random and independent transformations were applied to make a single random transformation because with only one, the new random

Table 2

Experiment 1	Experiment 2
DCGBF	ACGDE
DAGCI	BAGCF
BCGEF	DCGAF
HCGAF	DCGBF
DIGAF	CDGAF

1	2	1	2
DCAGF	EFGHI	DCAGF	EFGHI
R(1,2)=2.06		R(1,2)=4.55	

word often differed from the prototype in only one letter because only one of the letters in the input word was among the letters actually affected by the map.

The result finally was that the random word produced would typically have three or four letters in common with the original word. The letters in a word were left in the order resulting from applying the transformation.

Probability concordance was used to select b, more or less as in the previous example based on a search over b and CD. The set of elements S_i within a distance CD of each training word X_i were found. Then the average of $p(n, b, X_j) = \sum_i bK(bd(X_j, X_i))$ over the X_j in S_i is calculated. The correlation between these 5 averages and the corresponding frequencies in S_i is then found. Then the maximal correlation over the grid of values of CD and b is found and a value b for which this maximal correlation is obtained is determined.

The values of CD and b in the direct search were $CD = .5, 1.5, \ldots, 5.5$, $b = 1, 2, \ldots, 9$. In both experiments the optimal value of b was 3. Then finally $R_{n,b}$ for the test words 1 and 2, was calculated by using (1) in Section 3. This may be interpreted as an estimate of the ratio of the probability of finding a word near or equal to word 1 to the probability of finding a word near or equal to word 2.

To illustrate the psychological application imagine a subject is presented the "training words" and then two new words, it being understood that all the words were generated by some natural process. Then the subject is asked judge the relative likely hood of each of the new words 1 and 2 on the basis of this experience. In both experiments word 1 was the prototype word $DCAGF$ and word 2 was $EFGHI$. These two test words are shown in Table 1 just below the training sequences.

The kernel estimate of the probability ratio R for the two new words $1 = DCAGF$ v.s. $2 = EFGHI$, is called $R(1,2)$ and is computed from each of the two training sequences (with $b = 3$) and shown at the bottom of the two columns. In many similar test situations which were presented in classes as demonstrations of how the kernel estimate could be interpreted, the value of $R(1,2)$, as in this instance, appeared to conform qualitatively to the intuitive judgment of the relative odds of the two words by the subjects.

What this illustrates, apart from the methodology, is that our intuitive probability appears to be based on the "association by similarity" principle even though the long-standing idea of association by similarity is really different from that expressed by the kernel model. In the traditional formulation, it is "ideas" which are associated or "stimuli and responses." The assertion here is different in what may be a conceptually important way: it is probability of a future event that gets built up by experience. This hypothesis become clear only in the work of few psychologists, notably in the tradition of cognitive psychology as developed by E.C. Tolman(1932). The basic conceptual element in the work of Tolman, e.g., was called a "sign-gestalt expectancy" and was for all the world nothing more than a conditional probability, although the details of the development of such an "expectancy" were never worked out. The baysian tradition also models the development of subjective probabilities from experience. The kernel model is quite different and much closer to the psychological tradition of Tolman.

The similarity of the prototype word $DCABG$ with the words in the training sequence is clearly greater than the similarity of $EFGHI$ to these words. Also, the similarity measured by the number of common letters between $DCABG$ and $EFGHI$ is greater in Experiment 2 than in Experiment 1 which explains why $R(1,2) = 2.06$ in Experiment 1 and $R(1,2) = 4.55$ in Experiment 2. And these differences between the experiments are in the right direction, it appears. Although these results all conform to intuition in these and many other similar experiments which were done informally, only larger and more careful experiments with a wide variety of stimulus materials can show the usefulness of the kernel model for psychological purposes. It is nevertheless encouraging that these psychological predictions are obtained objectively and easily from the general kernel model.

The main alternative way of making such predictions at present is the neural net approach. This could be applied in this instance, and would no doubt yield similar results. But of course, this model and the kernel model for learning associations are not competitors. The neural net approach has potential value as leading to a neurological model.

APPENDIX.

Proof of Theorem 1. The pair (x, y) under consideration is always in M^* so

$$\lim_{r \to 0} \frac{P(S(x,r))}{P(S(y,r))} = R(x,y)$$

exists and is a finite positive number.

Lemma 1.

$$R(x,y) = \lim_{r \to 0} f(x,r)/f(y,r).$$

This is L'Hopital's rule applied to the definition of R as

$$\lim_{r \to 0} P(S(x,r))/P(S(y,r)),$$

considering $f(x,r)$ is the derivative with respect to r of $P(S(x,r))$. (See, e.g., Apostol, Vol. I, p. 394.)

Consider $p(b,x) = EbK(bd(x,X))$, the expectation being calculated using P. Observe there is a real random variable $D = d(x,X)$, with $P(D \leq r) = P(d(x,X) \leq r) = P(S(x,r))$. So regarded as a function of r, $P(S(x,r))$ is just the c.d.f. of D and its derivative $f(x,r)$ is therefore just the density of D. So we have

Lemma 2. $p(b,x) = \int_0^\infty bK(br)f(x,r)dr$.

Lemma 3. The limit as b goes to infinity of $p(b,x)/p(b,y)$ exists and is equal to $R(x,y)$.

Consider $p(b,x)/p(b,y) = \int_0^\infty bK(br)f(x,r)dr / \int_0^\infty bK(br)f(y,r)dr$. This is equal to

$$\int_0^\infty K(r)f(x,r/b)dr \Big/ \int_0^\infty K(r)f(y,r/b)dr$$

by a change of variable.

And from this there is a finite number c such that

$$(2) \qquad \int_0^c K(r)f(x,r/b)dr \Big/ \int_0^c K(r)f(y,r/b)dr \geq p(b,x)/p(b,y) - \varepsilon$$

for arbitrary $\varepsilon > 0$.

From Lemma 1 there is for an arbitrary $\delta > 0$, a number s such that $f(x,r)/f(y,r) \leq R(x,y)+\delta$ for $r \leq s$. Then if $b \geq c/s$, $f(x,r/b)/f(y,r/b) \leq R(x,y) + \delta$ for $r \leq c$, since $r/b \leq s$ in this case. So $f(x,r/b) \leq f(y,r/b)$ $(R(x,y) + \delta)$ for $r \leq c$. Thus using this inequality in the numerator of (2), valid over the range of integration, shows $R(x,y) + \delta \geq p(b,x)/p(b,y) - \varepsilon$. In other words for b greater than c/s, $R(x,y) + \varepsilon + \delta \geq p(b,x)/p(b,y)$. A similar argument gives $R(x,y) - \varepsilon - \delta \leq p(b,x)/p(b,y)$ for $b \geq c/s$. Since ε and d are arbitrary the proof of Lemma 3 is complete.

To complete the proof of Theorem 1 let ε be an arbitrary positive number. Note that by Lemma 3, there is a number b_1 such that

$$\left| \frac{p(b,x)}{p(b,y)} - R(x,y) \right| \leq \varepsilon \tag{3}$$

for $b > b_1$.

By the strong law, $p(n,b,x) \to p(b,x)$ and $p(n,b,y) \to p(b,y)$ a.s. as $n \to \infty$. And being strictly positive, since $f(x,r)$ is positive for r in a neighborhood of $r = 0$,

$$R_{n,b}(x,y) = \frac{p(n,b,x)}{p(n,b,y)} \to \frac{p(b,x)}{p(b,y)}$$

a.s. as $n \to \infty$ by an elementary argument. In other words, just interpreting the meaning of a.s. convergence, there is, for any b and arbitrary positive numbers δ_0 and γ, a number $n_1(\delta_0, b, \gamma)$ such that with probability at least $1 - \gamma$,

$$\left| \frac{p(n,b,x)}{p(n,b,y)} - \frac{p(b,x)}{p(b,y)} \right| < \delta_0 \tag{4}$$

for all $n \geq n_1(\delta_0, b, \gamma)$. Applying this with $b = b_1$ where b_1 is chosen as above to satisfy (3) and combining (3) with (4) shows that for all $n \geq n_1(\delta_0, b_1, \gamma)$,

$$\left| \frac{p(n,b_1,x)}{p(n,b_1),y)} - R(x,y) \right| \leq \delta_0 + \varepsilon \tag{5}$$

with probability at least $1 - \gamma$. Taking $\delta_0 + \varepsilon \leq \delta$, the numbers b_1 and n_1 required in Theorem 1 are provided and the proof is complete.

Proof of Theorem 2. From (1) of Section 3,

$$R_{n,b}(x_j, x_k) = \left(\sum_{i=1}^{n} K(bd(x_j, X_i))/n\right) \Big/ \left(\sum_{i=1}^{n} K(bd(x_k, X_i))/n\right),$$
$$\leq (f_j + (1 - f_j)K(bd_1))/(f_k + (1 - f_k)K(bd_2))$$

where f_j and f_k are, respectively, the relative frequency of $X_i = x_j$ and $X_i = x_k$, in the first n steps of the process, and d_1 is the smallest value of $d(x_j, X_i)$ among the X_i not equal to x_j, and d_2 is the largest value of $d(x_j, X_i)$ among the X_i not equal to x_k. The number b in front of K in both numerator and denominator has been canceled. Recall K is monotone decreasing.

Similarly we have

$$R_{n,b}(x_j, x_k) \geq \frac{(f_j + (1 - f_j)K(bd_2))}{(f_k + (1 - f_k)K(bd_1))}.$$

As n become large f_j and f_k approach p_j and p_k respectively by the strong law, and as b becomes large $K(bd_1)$ and $K(bd_2)$ both go to zero considering $\int_0^{\infty} K(r)dr = 1$ and that d_1 and d_2 are positive by hypothesis. This completes the proof.

REFERENCES

1. Apostol, Tom M.(1961), *Calculus*, New York, Blaisdell.
2. Davies,Roy O.(1971), *Measures not approximable or not specifiable by means of balls.* Mathematica. Vol. **18**, Part 2, No.36, December.
3. Rosenblatt, Murray (1971), *Curve estimates*, Ann. Math.Stat. Vol. **42**, No. 6, December, pp.1815-1842.
4. Thorndike, E.L (1911), *Animal intelligence: experimental studies*, New York:Macmillan.
5. Tolman, E.C.(1932), *Purposive behavior in animals and men*, NewYork: D.Appleton-Century Co.