

K-Optimal Randomization Tests for Association in Practical Metric Spaces Using Nearest Neighbor Methods

James MacQueen

University of California, Los Angeles

Jan Stallaert

University of Texas, Austin

Abstract

Let $X_i, Y_i, i = 1, 2, \dots, N$ be pairs of random variables, each X_i in M_1 , each Y_i in M_2 , where M_1 and M_2 are metric spaces with distances d_1 and d_2 , respectively. It is desired to test the hypothesis H_0 that the Y_i are identically distributed and independent of the X_i . Let S_i be the set of K nearest neighbors of X_i in distance d_1 and let V_i be the average of the distances $d_2(Y_j, Y_k)$ such that X_j and X_k are in S_i . Then $V = \frac{1}{N} \sum V_i$ will tend to be small if there is an association between the X_i and the Y_i and V can be used to test H_0 . An approximate randomization test based on the normal approximation to V was proposed by MacQueen (1991a). This was found to work in a wide range of situations but a satisfactory objective method for the choice of K was not established.

This paper provides a practical objective method of choosing K which is part of a new test. The test statistic p^* is defined as the minimum over $K = 2, 3, \dots, N - 1$ of the significance probabilities of the above test. This test statistic being generally too difficult to obtain exactly, is evaluated approximately by taking the minimum over the normal approximation estimates \hat{p}_K of the significance probabilities p_K for each K . The resulting test statistic $\hat{p}^* = \min_K \hat{p}_K$ is then evaluated by approximate randomization based on a sample of random pairings of the X_i and the Y_i , getting \hat{p}^* for each and then calculating an approximate significance probability \hat{p} as the proportion of these less than or equal to \hat{p}^* from the original data. Because \hat{p} is an unbiased estimate of the true significance probability and can be made as accurate as desired by increasing the number of random pairings, it can be used as a measure of strength in the usual way, and if a formal procedure is desired, rejecting H_0 if $\hat{p} \leq \alpha$ will accomplish this with a Type I error of not more than $\alpha + \varepsilon$, with ε small.

The test is applied to a variety of simulated data sets of quite different kinds and found to be a practical and convincing test. In the multivariate situation the test performs respectably well in comparison to the F test when all the assumptions of this test are in effect.

Keywords: Association; Metric Space; Randomization Test

1 Introduction and Overview

Given a sample of N pairs, $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$, X_i in M_1 , Y_i in M_2 , where M_1 and M_2 are metric spaces with respective distance functions d_1 and d_2 , it is desired to test the hypothesis H_0 that the Y_i are independent and identically distributed and also independent of the X_i . This paper describes a test which is intended to work against the large class of alternatives to H_0 , where for each x in M_1 there is a distribution for Y given $X = x$, and this distribution tends to concentrate in a region which varies smoothly with x .

A simple test of H_0 based on K nearest neighbor regression (Fix and Hodges, 1951, 1952) has been proposed by MacQueen (1991a). This test is as follows: Let S_i be the set of K nearest neighbors of X_i in distance d_1 , and let T_i be the corresponding set of Y values. Let V_i , here called the *within-set variation* of the Y_j in T_i , be the sum of the $K(K-1)/2$ pairwise distances between the Y_j in T_i , divided by $K(K-1)/2$. Then $V = \frac{1}{N} \sum_i V_i$, which can be called the *average within-set variation*, will be relatively small under the above described alternative to H_0 . With this test statistic V , a formal randomization test with Type I error α is easily performed. This is done by finding a largest number V_α such that $P[V' \leq V_\alpha] \leq \alpha$ where V' is computed from a random pairing of the sample Y_i with the X_i . Then H_0 is rejected if $V \leq V_\alpha$. Alternatively, the significance probability of the average within-set variation V , $p_K = p_K(Y) = P[V' \leq V]$, may be calculated and H_0 rejected if $p_K \leq \alpha$.

This test is an exact conditional test, that is, given the data, with H_0 holding, the Y_i are already paired at random with the X_i and so the distribution of V under H_0 is equal to the randomization distribution of V , which in principle we know from the data.

The probability p_K can be computed exactly in very small samples, $N \leq 12$, say, and approximated very well by calculating V' for a large sample of random pairings. It can also be estimated reasonably well by a normal approximation to the distribution of V' . This latter approximation has an important role in implementing the K -Optimal test.

Unfortunately, the performance of the above test depends strongly on K and no objective basis for choosing K is immediately evident. The “ K -Optimal” test of H_0 was devised

as a solution to this problem. Briefly, the “ K -Optimal” test is based on (approximately) minimizing the significance probability p_K for the above test by a direct search over K to produce an “Optimal” K called K^* . As explained in Section 2, randomization is used to control Type I error. The main purpose of this paper is to describe this latter test and report some simulation experiments illustrating and substantiating its usefulness.

Contrary to most conventional methods for detecting association, the proposed method does not require an explicit estimate of a regression function to calculate the significance probability. The test is based directly on the data and the logic underlying the nearest neighbor idea as expressed in the alternative to H_0 described above. This itself may be regarded as a paraphrase of a fundamental idea in science called, usually, the continuity of nature and expressed, often, by saying ‘like causes have like effects’. Here this becomes the simple idea that if sample elements in the X space are close, the corresponding elements in the Y space should tend to be close. (But note another kind of departure from H_0 described and illustrated in Section 3.2 below for which the K -optimal test is automatically sensitive, and thus provides for sensitivity to a kind of relationship not afforded by any conventional test for association known to the authors.)

However, if the final goal is to develop a predictive model, the proposed method can still be useful as a preliminary step, and if H_0 is clearly rejected one can at this point develop a suitable predictive model. In so far as the proposed method has good power against a wide variety of alternatives, it reduces the chance of adopting an erroneous attitude towards H_0 in the direction of accepting H_0 due to an inappropriate predictive model, e.g., using a linear regression function instead of some non-linear one. For this reason it was deemed imperative to get a reading on the power of the test in a context where a known and good method was available for comparison.

In the experiments reported in section 3, we compared the new method with methods specifically designed for the situation being examined. In the case of Euclidean data — with both linear and non-linear regression functions — the comparison was made with the F test, which is known to give good, if not maximal, power, and from the tables in section 3 it can

be seen that even in these situations the K -Optimal test gives results very close to the ones obtained with these specialized methods. We did not specifically compare the new method to non-parametric regression methods. The reason is that the multitude of such methods would — practically speaking — yield an unwieldy computational task outside the scope of this paper, lest one restricts his/her attention to an arbitrary selection of them. But, as mentioned before, all regression methods (e.g., ordinary linear regression) detect association when, given the observed X values, the predicted values and observed values for Y are close, and they proceed by rejecting H_0 if the total of these residuals (or their squares) is small. Our method does not go through this intermediate prediction step and detects association based directly on the observed Y values. In our opinion, the fact that the proposed method does not depend upon the power of any predictive method is a useful advantage, for it means that the method can be used to complement the other (prediction-based) methods to detect association. The method is perhaps maximally free of parametric assumptions as among such tests, and that it uses only distance (and/or similarity) measurement strongly suggests it will have equally good power in the nominal domain of application where such measurement is all that is available.

A considerable use is made here of $\sum_{ij} d_{ij}$ as a measure of sample variation for metric space data. Cover (1968, p.52) introduced and defined the idea of the r th moment of distribution P on a metric space, by $Ed^r(X, X')$ where X and X' are i.i.d., with the common distribution P . This is the same idea as “variation” whether applied to a sample or the population, but we prefer the term variation on the grounds that the term moment used with numerical variables almost never measures variability or variance even. (For this, the term central moment has to be used. This concept itself can be applied to metric space variables by taking the average distance from a “center” or “centroid”. Sverdrup-Thygeson (1981) has treated this idea and established the law of large numbers for the centers and the variation around them.)

Whatever it is called, the concept of variation seems to be a fundamental idea for statistics in metric spaces. At the same time it is a relatively simple and easily understood idea, even

simpler than the idea of “sample variance” $\frac{1}{N} \sum (x_i - \bar{x})^2$ used with numerical idea. The latter is in fact 1/2 the sample variation of order 2 when the space is actually Euclidean, and so, except for factor 1/2, “variance” is a special case of variation. For further remarks on the simplicity of variation, which has some pedagogical significance, see Section 4 below. The formula for the variance of the sample variation of order r is given by MacQueen (1991b).

Cover’s remarks suggest a general way of measuring association for the joint distribution of Y and X , and/or departures from H_0 . This is to calculate at each x , Cover’s second moment of Y given $X = x$, that is $E(d^r(Y, Y') | X = x)$ defined as above except now Y and Y' are independent with the conditional distribution of Y given $X = x$. Call this $V^r(Y | X = x)$. If Y and X are independent then this will be equal to $V^r(Y)$ and some reasonable and convenient measure of the difference between $V^r(Y | X = x)$ and the constant $V^r(Y)$, such as the expected squared difference, becomes a measure of departure from independence.

In these terms the proposed test may be described as being sensitive to departures from H_0 of the form $E(V^1(Y | X)) < V^1(Y)$. It is to be noted that $V^1(Y | X = x)$ might be larger than $V^1(Y)$ for some x regions but smaller in others, and average to something close to $V^1(Y)$ so the proposed test is not suited for detecting this kind of relationship.

While it appears to be possible to devise a test which would be sensitive to the more general class of alternatives, that is, $V^r(Y | X = x) \neq V^r(Y)$, a really good way to proceed is not evident. One possible test might be based on the V_i , since each is a plausible estimate of $V^1(Y | X = X_i)$ in lieu of a sample where there are a good number of Y ’s for each different X . Then, for example, $T = \sum (V_i - V^*)^2$ where $V^* = \sum d(Y_i, Y_j) / (N(N - 1))$, the sum being taken over all pairs i, j , would provide a test statistic which should be sensitive to “almost any” departure from H_0 .

However, implementing such a test still leaves the problem of choosing K open, and implementing the obvious “ K -optimal” variant would present formidable problems. Also in the design of the K -Optimal test, the alternative to H_0 of primary interest is where the distribution of Y given $X = x$ is concentrated in some region which varies smoothly with x . There is no obvious reason why a test based on T would be any improvement in this case.

So this range of possibilities is left open for future research on the problem of detecting a relationship in which the variation of the distribution changes with X , but where the location of the distribution tends to be constant.

The notion of a metric space is very general, of course, so the potential range of application of tests for association with metric space data is very great. All that is required is that there be a relevant way of measuring the distance between pairs of elements of the kind of interest. If all else fails, this is usually possible to do by judgment methods, making available tests for association in certain complex situations arising in the social sciences, as, for example, the relationship between the “form” of government and the amount and “form” of war, or between the “management style” and the productivity of the firm. In these cases, similarity is often the more natural concept to use, as opposed to dissimilarity or “distance”, but such data is easily handled by a transformation such as $s = \exp(-d)$ and its inverse, d being distance, and s being similarity on a scale where $s = 1$ means identity and $s = 0$ is the maximum possible dissimilarity. Also, it should be noted that if sample elements are described in several different metric spaces — as is the case in ordinary multivariate data — they can be combined into a single metric space by taking a weighted average of their respective distances.

It is this wide range of potential applications that is interesting to the statistician and accounts for the term *practical* in the above title, which is to be emphasized, lest the reader thinks our interest lies only in the mathematical elegance of the metric space idea.

While there is a substantial literature on K nearest neighbor methods for multi-variate data, there is only a relatively small amount of work on the association testing problem at the general metric space level. Fix and Hodges (1951,1953) introduced nearest neighbor regression ideas and applied them to the classification problem. Cover and Hart (1967) give a general formulation of the nearest neighbor concepts in a general metric space context. They formulate the regression problem itself in a very general way. Specifically, they introduced a general decision problem with sample X 's in a metric space, which is that of estimating a parameter θ associated with each X , with this parameter itself residing in a general metric

space, and this decision problem has an arbitrary loss function. This model contains the model used here in which both X_i and Y_i are in metric spaces although Cover and Hart do not treat the test for the association problem per se. MacQueen (1968) discussed the possibility of testing for association using clustering. This method is described in Section 4 below and compared with the proposed K -Optimal test, which is found to be superior.

A good portion of the work on statistics in metric spaces, including hypothesis testing by randomization, has been reviewed by Diaconis (1988) who gives a number of references. Good (1994) reviews many applications of randomization tests including several which make use of a general distance function although none of the findings he describes are directly comparable to what is done here.

The method described here only tests whether there is some association between X and Y and we assume the X space is fixed, even though it may itself be a product space. The approach is easily extended to the *variable selection* problem using one or another of the familiar approaches to this problem, for example, the well known non-parametric approach of Forsythe et al. (1973). Following their logic, the K -optimal test would be applied to the original data with and without a variable included, using the weighted sum metric on the product spaces as suggested above, with the coefficient adjusted to produce unit variation for each X variable. The difference between these two significance probabilities would become a test statistic, and the final tests would be obtained by repeating this on many samples where the X variable to be added was put in a random order with respect to the other variables. The final significance probability for adding the variable would be based on randomization distributions of these differences, and would consist of the proportion of difference in the randomized sample at or below the difference from the original data. If significant, the claim would be that the improvement in the test statistic in the original data was too great to be accounted for under the hypothesis that the variable in question was unrelated, even jointly, with the other variables. The final significance probability itself would be a useful index of strength of the variable and might be used for comparison with other variables that might be used. Of course, if an irrelevant variable were to be added, we expect the significance

probability to go down.

This is getting a bit computational intensive, but is easily seen to be feasible even on the PC, with moderate sample sizes, and modest number of variables, for say $N = 100$ and seven predictor variables.

We have run the K -Optimal test on many simulated data sets of the kind found in the statistical literature. So far, the K -Optimal tests has always done about as well as other methods which it occurred to us to try, and sometimes much better. The experiments reported here are typical of those we have done and are chosen to illustrate the general tenor of our experience.

The paper is organized as follows. The K -Optimal test is described in more detail in Section 2 and its practical implementation is described in a subsection there.

Section 3 presents a variety of applications of the test to simulated data sets in order to illustrate the test and to help evaluate its usefulness.

In these experiments, significance probability is interpreted somewhat freely as an indication of statistical power, which is of course the real objective of the research. If test A has a smaller significance probability than test B on the same data set, referred to conventionally and somewhat imprecisely by saying the test A is “more significant”, then it is reasonable to conclude that for the alternative to H_0 at hand — vague though it may be — the probability of rejection for test A is obviously larger than for test B, for the entire range of prespecified α levels between the two significance probabilities. In this case there is an instance — a sample of size one, if you will — where test A has superior power to B, for a range of α values, and for the general kind of alternative hypothesis described above.

Although in any given situation a more extensive power study would be possible and power curves could be obtained, we were content to sample the power using just this significance probability comparison, but do so for a relatively wide range of situations. The entire range of possible situations is hopelessly large and no attempt to sample this range systematically was made. Instead, the situations were chosen on the basis of interest on the part of the investigators, but hopefully with some appeal to the interests of scientists and

statisticians generally.

Section 3.1 gives a sample of comparisons of the K -Optimal test to the standard multivariate test for association with multivariate data where a conventional test based on F is available for comparison. It is found in these comparisons that the K -Optimal test does almost as well as the standard test even when all the assumption of the standard method are in place (i.e., normally distributed errors with constant variance and a linear regression function).

It was expected on rough intuitive grounds that, if the relationship between X and Y is relatively “noisy” — meaning the sample Y have high within-set variation for each X — then K^* would be relatively large, (other things being equal) because a larger K would mean each of the within-set variations V_i and hence V would have relatively low sample variation, which as usual, should translate into stronger significance and so a larger K would be produced from the search. To check this intuition σ_ϵ was varied systematically in some of the samples described in Section 3.1. The value of K^* did tend to increase with σ_ϵ in confirmation of this.

On the other hand, it is intuitively clear that relatively small values of K have the potential advantage being able to track the relationship between X and Y more precisely and hence produce lower and possibly more significant values of the test statistic V , which is chosen just to be able to do this. This suggests that if the underlying relationship between X and Y is relatively complex, meaning the location of the Y 's tends to vary rapidly with X , the values of K^* would tend to be relatively smaller. To check this hypothesis, an experiment was done in which the “noise” was kept constant, while the complexity was varied. Specifically, a sinusoidal regression function was used, and the frequency was stepwise increased. The results from this experiment are described in Section 3.2. The value of K^* was found to decrease as the frequency increased, as expected. The potential sensitivity of K^* to these important features of the underlying relationship and the possibility that it makes a good tradeoff between these conflicting factors, is regarded as a promising feature of the method.

Section 3.2 also discusses briefly the possible interpretation of a *much larger-than-expected*

value for V . This can happen if pairs of X points which are very close tend to have their associated Y values further apart than one would expect under H_0 . Application of the K -Optimal method to measure and detect this is straight forward and may be of interest in certain situations, and a simulated data set is analysed to illustrate the possibility.

Section 3.3 shows how the test may be applied in the commonly occurring situation where the observed elements are described by lists of qualitative attributes. It also illustrates how with such relatively complex and qualitative data, the test might show clearly the presence of a strong relationship which would not be obvious from inspection. For this data set the K -Optimal test is compared with another promising method based on clustering (MacQueen, 1965, 1968), and is found to be noticeably more powerful.

Section 3.4 illustrates application to 0/1 data where the true regression function is a Boolean function of k variables, observed with error in the dependent variable. It turns out to be surprisingly easy to detect the presence of such highly non-linear relationships with the K -Optimal tests, and the reason (somewhat obvious in retrospect) appears to be that “simple” Boolean functions have a certain kind of limited continuity. Specifically, the value of such functions is usually constant over small subsets of adjacent values of the predictor variables. The K -Optimal method locates the right and relatively small values of K where the dependent values are almost constant and the test statistic obligingly has a relatively small value for this reason.

Section 3.5 offers a brief discussion of alternative methods of choosing K , including several rules of thumb for choosing K , and also a method based on optimizing V instead of p_K . It is argued that, in general, the K -Optimal method is superior to the other methods.

2 The K -Optimal Test

The sets of nearest neighbors require a convention in their definition because of ties. Here S_i , the set of K nearest neighbors of X_i , is determined by first ordering the X 's in increasing d_1 distance from X_i , taking any tied values in increasing order of their subscripts. Then S_i is just the first K elements in this order.

For simplicity in notation, abbreviate the sample sequence X_1, X_2, \dots, X_N by X and the sample sequence Y_1, Y_2, \dots, Y_N by Y . Then define V just as above, that is, $V = V(Y) = \frac{1}{N} \sum_i V_i$ where V_i is the sum of all pairwise distances $d_2(Y_j, Y_k)$, divided by $K(K-1)/2$, such that X_j and X_k are in S_i . Then letting Y' be a random permutation of the elements of Y , define p_K , as above, by $p_K(X, Y) = P[V(Y') \leq V(Y)]$ or just $P[V' \leq V]$. As was mentioned above, p_K is the conventional significance probability of the randomization test where H_0 is rejected if V is less than or equal to a certain critical value, V_α , this being chosen so that the probability of rejection is at most α for some specified α , e.g., 0.05. Of course, p_K may itself be regarded as a test statistic and the test is to reject H_0 if $p_K \leq \alpha$.

Now consider a different test, whose test statistic is $p^* = p^*(X, Y) = \min_K p_K(X, Y)$ where the minimum is taken over the meaningful choices of K , that is, the range from $K = 2$ to $K = N - 1$. With p^* so defined, let $p = P[p^*(X, Y') \leq p^*(X, Y)]$ where again Y' is a random permutation of the Y_i . The “ K -optimal” test with Type I error α , is to reject H_0 if $p \leq \alpha$, and p is its conventional significance probability.

It is clear after a little consideration that p is too difficult to obtain exactly in most cases, and we will be content to work with the approximation defined in Section 2 just below.

Notice that comparison of different values of K through p_K , rather than through the test statistic V is critical to the idea of K -optimal tests, since the meaning of V changes as K varies, depending heavily on the spread of the distribution of V . The different values of V as K varies become directly comparable through transformation to p_K . Note also that even though p is a complicated function of the data, it has a conceptually straight forward randomization distribution and the Type I error of the procedure of rejecting H_0 if $p \leq \alpha$ is exactly α . Thus in spite of the fact that p^* is based on a search over the $N - 2$ meaningful values of values of K , the test is not biased in the direction of excessive rejection of H_0 . The search for the optimal K represents a certain amount of data “massage” but since the “massage” is applied equally to the randomized data sets, it has been rendered harmless.

The main risk is bias in the other direction, that is, towards low power. The search for the minimum over K of $p_K(X, Y')$ may take advantage too well of the appearance of

relationship occurring in the pairing of random permutations Y' with X so that $p^*(X, Y')$ will too often be small and below $p^*(X, Y)$ thus raising p more than we would like. Thus, while under H_0 , the Type I error is exactly α , the critical question is, does the test have a useful level of power? The experiments reported below answer this affirmatively.

Note that p , being the conventional significance probability for the test, could be employed as usual, e.g., in comparing the power of different methods or in using p values as indices of the relative strength of certain relationships, and generally, using the significance probability as an aid to judgment. The estimate of p described below may be used in this same way.

Implementation

To implement the K -optimal tests two approximations are used. First, the normal approximation is used for each individual $K = 2, 3, \dots, N - 1$, to get an estimate \hat{p}_K of each p_K , and then $p^* = \min_K p_K(X, Y)$ is approximated by $\hat{p}^*(X, Y) = \min_K \hat{p}_K(X, Y)$. Second, the significance probability $p = P[p^*(X, Y') \leq p^*(X, Y)]$ is approximated by what will be called here “brute force” randomization: A random sample of R permutations, Y'_r , $r = 1, 2, \dots, R$, is taken using the computer. For each $r = 1, 2, \dots, R$, $\hat{p}^*(X, Y'_r)$ is computed as just described, that is, by applying the normal approximation for each $K = 2, 3, \dots, N - 1$ to get an estimate $\hat{p}_K(X, Y'_r)$ of the probability that (another) random permutation, say Y'' , would have $V(Y'') \leq V(Y'_r)$, and then taking the minimum of these over K to get $\hat{p}^*(X, Y'_r)$. Finally the significance probability $p = P[p^*(X, Y') \leq p^*(X, Y)]$ is estimated as

$$\hat{p} = \frac{\text{Number of } r \text{ such that } \hat{p}^*(X, Y'_r) \leq \hat{p}^*(X, Y)}{R}$$

Actually, since the normal approximation $z = (V' - EV')/\sigma(V')$ is monotone with \hat{p}_K , the search over the K is done with the corresponding z 's and in fact the relative frequency is likewise just the number of times in the sample of R permutations Y'_r , the lowest z over the $N - 2$ values of K falls at or below the lowest z value for the original sample. This minimal z -value is called Z^* .

The formulae for EV' and $\sigma(V')$ have been derived in a slightly different form by Mac-

Queen (1991a). They are given in Appendix 1. (A transcription error in this reference is corrected.)

Note that \hat{p} is the unbiased estimate from the sample of R values, of the true significance probability of the test statistic \hat{p}^* regarded as a test statistic in its own right, and this is true even if there is error in taking the minimum value over the \hat{p}_K instead of the p_K . For this reason, we will hereafter use the term *K-Optimal* to refer to the test based on \hat{p}^* and then \hat{p} is an estimate of the significance probability. Nevertheless, for understanding the method, and for theoretical reasons, it is desirable to know something about the accuracy of this step in the approximation.

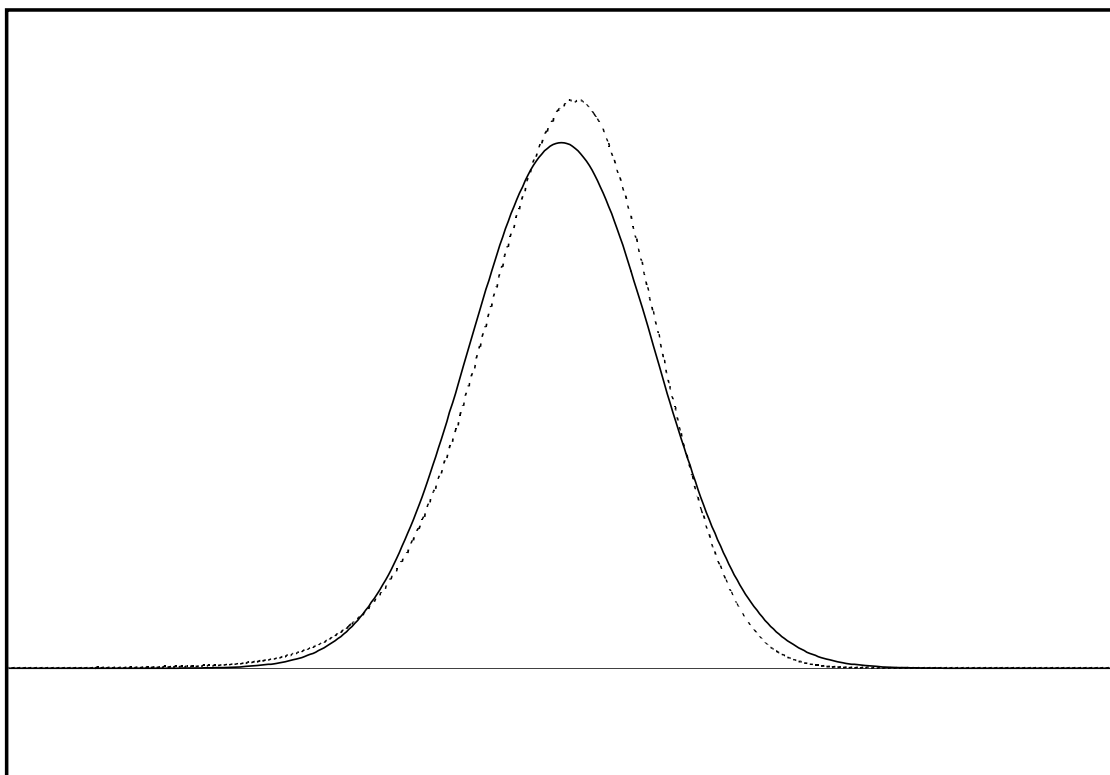


Figure 1: Empirical Distribution of V' and Normal Approximation

This was checked directly by simulation for a few data sets with $N = 30$. That is, at

each K , p_K was estimated directly by 1,000,000 randomizations. A small systematic bias in the normal approximation was discovered, and this is illustrated in Fig. 1, where a plot of the normal approximation based on the exact formulae for the mean and variance of V' is compared with the plot of the empirical distribution of V' based on sampling 1,000,000 random pairings. The solid line shows the normal approximation, the dashed line the empirical distribution of V' .

The bias shown here was very consistent as K varied. Because the slight bias was consistent in shape, the order of the significance probabilities for the two methods as K varied were virtually identical. Thus the optimal K itself as between the two approximations, differed only by one or two in nearly every case, and of course the z values and the estimated significance probabilities differed but little between the two methods. This was especially true when the relationship in the data was very sensitive to K . On the other hand when the data was actually linear, there was a tendency for the p_K values to vary only slightly over a wide range of K values, and so of course the optimal K would be more variable. But because the p_K varied only slightly even if the values of K for the two methods differed, final estimates were virtually identical for the two ways of estimating the p_K .

In any case, our conclusion is that it would not make much difference whether a large sample of permutations was used at each K to get an estimate for p_K , or the normal approximation, but the latter is preferable simply because it is free of sampling error itself, and taking very large samples at each K is impractical, considering that it is necessary to take such a sample at each K and then do this for each r in the brute force simulation.

From Figure 1 it can be seen that in the left tail the normal approximation *underestimates* $P[V' \leq V]$ and if one were to base his judgment on the significance probability obtained by such approximation, a decision in favor of the alternative to H_0 will too often be adopted for a range of small α values. For example, for a data set with $N = 30$ and $K = 8$ (see the data set lin1 in Section 3.5), a z -value of -2.67 was calculated, but a brute force randomization with $R = 100,000$ shows that an estimate for $P[V' \leq V]$ is 0.0160 and it is difficult to reconcile this probability estimate with $P[Z \leq -2.67] = 0.0038$ obtained from the normal

distribution.

Now, even when \hat{p}_K underestimates p_K in some range, our estimate \hat{p} does not systematically underestimate p in any range, since in the K -Optimal test the values for \hat{p}_K lose their interpretation of probability and are merely regarded as test statistics, and the significance probability p is directly estimated by simulating its actual distribution via brute force randomization. Unlike the normal approximation for the individual values of K , we were not able to find an accurate and easily implementable mathematical approximation for the distribution of p^* , the *minimum* of the $N - 2$ z -values.

The error produced by random sampling from the permutation distribution in the “brute force” randomization, is, practically speaking, a negligible problem in most cases, because in today’s computational environment, R can usually be made as large as necessary to ease any doubts about the statistical significance revealed by \hat{p} .

A detailed analysis of the sampling error in approximate randomization tests based on sampling from the permutation distribution has been given by Marriott (1979). The point is that the number of times $\hat{p}^*(X, Y') \leq \hat{p}^*(X, Y)$ is a binomial random variable with the parameter $P[\hat{p}^*(X, Y') \leq \hat{p}^*(X, Y)]$ and if R is small and cannot be increased, a more refined analysis of the meaning of \hat{p} can be made on this basis. For example, there are confidence limits for the true significance probability and the standard error of \hat{p} is estimated as $s_{\hat{p}} = \sqrt{\hat{p}(1 - \hat{p})/(R - 1)}$. These refinements do not appear to be of much practical interest in the situation under consideration because it is usually possible to make R very large if necessary. In our test results, we used $R = 2,000$, unless otherwise stated, so for $\hat{p} = 0.01$, $s_{\hat{p}} = 0.002$, and a little consideration shows that treating $\hat{p} = 0.01$ as if $p = 0.01$ is not a serious error.

3 Illustrative Applications

3.1 Application to multiple regression

Here the multiple regression situation is chosen for comparison in order to obtain a reading on how well the K -optimal tests perform in a situation where there is a known test with good if not maximal power. A number of data sets were simulated with linear and non-linear regression functions, with approximately normal errors having constant variance based on the sum of 12 computer generated numbers, uniformly distributed on $(0, 1)$. The conventional F test based on linear regression was applied to these sets, along with the K -optimal test, using Euclidean distance for the X vectors. Samples of various sizes were generated, also varying in the number of predictor variables, and in strength of association. Also, several different polynomial forms were generated.

Problem	k	N	K^*	Z^*	σ_ϵ	Obs R^2	$P(F)$	$P(R^2)$	\hat{p}
lin0	1	30	16	-2.4578	3	0.11	0.0398	0.0408	0.1155
lin1	1	30	16	-3.7974	2	0.24	0.0034	0.0035	0.0095
lin2	1	30	14	-4.8023	1.5	0.35	0.0003	0.0003	0.0005
lin3	1	30	14	-8.3370	1	0.57	0.0000	0.0000	0.0000
lin4	3	30	14	-3.0941	5	0.20	0.0314	0.0327	0.0150
lin5	3	30	6	-3.3562	3	0.46	0.0003	0.0003	0.0075
lin6	5	30	3	-1.2300	8	0.00	0.4475	0.4451	0.4335
lin7	5	30	3	-1.4906	4	0.10	0.1847	0.1843	0.3225
lin8	5	30	3	-1.7240	3	0.17	0.0858	0.0868	0.2275

Table 1: Linear Regression with $f(x) = \sum_{i=1}^k x_i$

Some typical results for the simple linear regression are given in Table 1. There were nine simulated data sets as indicated in the table, chosen to explore the range of low to moderate R^2 values. These varied in the number of predictor variables as well, as given in the column labeled k . In each case the X 's were independent and uniformly distributed. The unadjusted R^2 from ordinary multiple regression is given for each sample in the column labeled accordingly. The usual significance probability based on the F test is in the column labeled $P(F)$. In addition, an approximate randomization test was applied to R^2 itself. This was based on sampling from the distribution of R^2 under random pairing of the Y values

with the X vectors. For each model, for each sample, 100,000 randomizations were taken and the results are displayed under the column labeled $P(R^2)$. Comparison of this column with $P(F)$ confirms that the F test is robust and is a credible estimate of the true significance probability estimated under H_0 , a fact that is generally accepted in practice.

As expected, the F test gave stronger significance levels in most data sets. The data set lin0 gave a significance probability of 0.0398 using the F test, but only 0.1155 for the K -optimal method. However this is the only instance where the decision would be different under the conventional 0.05 level test or where the overall judgment based on conventional interpretations of significance probabilities might be different in a substantive way. A somewhat similar discrepancy is seen in the set lin8. The other data sets would lead to virtually the same decision as a practical matter for both tests with a linear or approximately linear regression function and with roughly normal errors.

The conclusion is that in this domain the K -optimal test performed quite well in comparison to the more powerful F -test. Considering that the actual data is probably never really linear, and the errors are usually not really normal, the K -optimal test is probably preferable as a method for detection of association in this range of data.

Table 1 also shows evidence on the relation between σ_ϵ and K^* suggested in the introduction. The four samples lin0 through lin3 are generated exactly in the same way except σ_ϵ is decreasing from 3 to 1. Note that K^* also decreases roughly. This is also true for lin4 and 5 which are identical, with K^* decreasing from 14 to 6 as σ_ϵ decreases from 5 to 3. The samples lin6, lin7 and lin8 show only that K^* is not increasing, possibly because K^* is nearing the minimal value of 2.

The non-linear data sets for which multiple regression was applied are shown in Table 2. In these data sets the true R^2 was chosen to be larger than in the linear data sets in order to have a comparison in this range. But the multiple regression method was given the strong advantage of having the correct nonlinear function (see Table 2), so only the coefficients and the constant term of the three non-linear components of the regression had to be estimated. With this considerable advantage the linear method is surely an optimal or nearly optimal

Problem	σ_ϵ	Obs R^2	$P(F)$	\hat{p}
s1n30std2	2	0.94	0.0000	0.0000
s2n30std3	3	0.87	0.0000	0.0000
s3n30std5	5	0.68	0.0000	0.0020
s4n30std7	7	0.50	0.0001	0.0240
s5n30std10	10	0.30	0.0064	0.2210
s6n30std5	5	0.53	0.0000	0.0000

Table 2: Non-linear Data sets with $f(x) = x_1x_2x_3 + x_1 \sin x_2 + x_3^2$

procedure. Nevertheless the K -Optimal test found some evidence for association in five out of the six examples (using the conventional 0.05 criterion). Of course application of the linear regression itself without using the non-linear terms was expected to produce essentially a horizontal regression surface and did so in the few instances which were actually computed. A hunting expedition with trial an error over polynomial forms would probably pick up good evidence of association here. But how would one proceed with such a search in the range of highly nonlinear data and how much adjustment should be made in the significance level reported by the F test when, say, the three squared terms were added? So here, where the chances of model misspecification are very high the K -Optimal method may be preferred because of the objectivity and accuracy and distribution-free character of the significance probability it provides.

3.2 K -optimal tests as a function of complexity

To examine the question of how K^* varied as a function of the complexity of the underlying regression function, five sinusoidal regression functions on the interval $[0, 1]$ were used. These were $f_j(x) = \sin(2\pi\nu_j x)$ where the frequency ν_j was varied in steps from 0.25 to 3. Samples of 30 pairs were prepared at each of these frequencies with X having values at 30 equally spaced values, that is, at $X_i = (i - 1)/29$, $i = 1, 2, \dots, 30$, with $Y_i = f_j(X_i) + \epsilon_i$ where the ϵ_i are i.i.d., and approximately normal with a standard deviation of $\sigma_\epsilon = .3$. The data set sin300 is plotted in Figure 2 as an illustration.

The results are presented in Table 3. A point of interest is that the value of K^* tends to

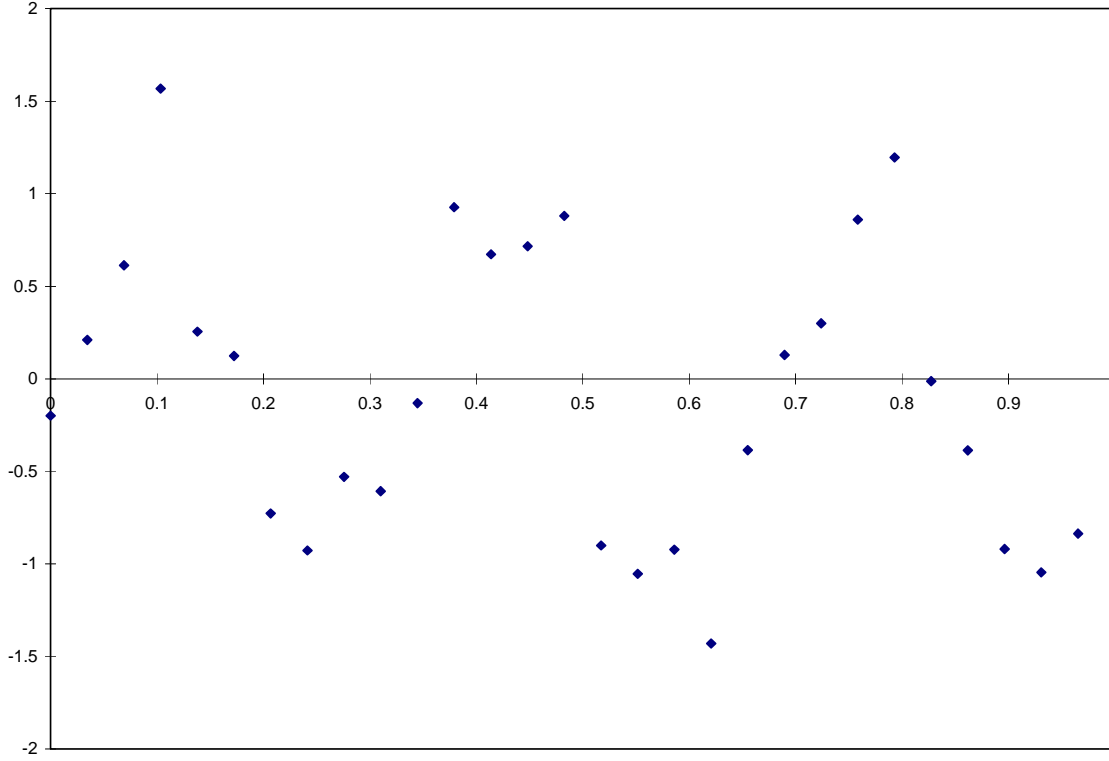


Figure 2: Scatterplot of the sample for `sin300`

Problem	ν_j	n	K^*	Z^*	σ_ϵ	$P(d')$	\hat{p}
<code>sin25</code>	0.25	30	11	-5.1690	0.3	0.0135	0.0005
<code>sin50</code>	0.50	30	11	-3.0488	0.3	0.0305	0.0300
<code>sin100</code>	1.00	30	7	-9.1651	0.3	0.0000	0.0000
<code>sin150</code>	1.50	30	5	-6.3559	0.3	0.0000	0.0000
<code>sin200</code>	2.00	30	3	-4.8653	0.3	0.0000	0.0005
<code>sin250</code>	2.50	30	3	-4.8403	0.3	0.0000	0.0005
<code>sin300</code>	3.00	30	2	-3.7442	0.3	0.0001	0.0075

Table 3: Sinusoidal functions with increasing complexity

decrease as the frequency increases. The smallest value of $K^* = 2$ occurs with sin300 which has the highest frequency. Thus the regression function goes through three full cycles over the range of the X 's.

This table includes another test for association based on the Durbin-Watson statistic. This statistic is commonly and successfully used in time series to see if there is any time dependent relationship in the residuals from a time series regression analysis. Using this approach, we found that — except for the data set sin50 — the Durbin-Watson test found positive auto-correlation at the 1% level, based upon the critical values in the Savin-White tables. So, the results of the latter test are consistent with the results using the K -Optimal method. But, in order to get a better estimate of the significance probability for rejecting H_0 , we modified the original Durbin-Watson test statistic, and this test statistic d' was applied directly to the Y values rather than to the residuals, and was evaluated directly by sampling the randomization distribution. Thus $d'(Y) = \sum_{i=2}^n (Y_i - Y_{i-1})^2$ was computed for each of the samples in Table 3, and also for 100,000 random pairings of the Y 's with the X 's for each of the six samples. The column labeled $P(d')$ is the proportion of the random pairings with $d'(Y')$ at or below the observed $d'(Y)$ in each sample and is a good estimate of the significance probability of d' in each case.

The K -Optimal significance probability in the column \hat{p} (only 2000 randomizations) is essentially the same as that of the Durbin-Watson statistic except in the case of sin25. Inspection of this data indicate the difference is probably due to two large successive differences whose effect on the Durbin-Watson test is enhanced by squaring. Using $\sum |Y_i - Y_{i-1}|$ in place of $d' = \sum (Y_i - Y_{i-1})^2$ as a test statistic would probably make the results from the two methods even more similar.

That the K -Optimal performs as well as the randomization test using d' — a test statistic designed specifically for this situation — is encouraging all though not surprising. If K had been set at 2 and V used as a test statistic without a search over K , the two tests would be virtually identical because of the convention of taking ties in order of subscript. This means the two points in each set of nearest neighbors would be the same as those taken in

the modified Durbin-Watson statistic, except at the end points.

Similarity of \hat{p} and $P(d')$ bears also on the question raised in Section 2 of possible loss of power for the K -Optimal test because the search over K would find among the randomizations too many low z values, just by chance alone. The advantage of the search over K for the original data tends to compensate or more than compensate, it is hoped, for the former search. It appears that in this instance the compensation is almost complete, and accounts for the fact that \hat{p} shows as much strength of association as $P(d')$ even though d' is a test statistic designed for this situation.

As indicated in the introduction, inspection of the test statistic V shows that if the Y values associated with several X values are very far apart, more so than would be expected if the Y 's are independent of the X 's, then z would tend to be positive. Thus to detect this kind of very complex relationship — bordering on discontinuity — the search over K would be made for the maximum z , and the estimated significance probability would be defined accordingly as the proportion of random pairings where z^* , the maximal z , was as large or larger than z^* from the original data. Once again, the Type I error would be distribution free.

To illustrate this possibility, the above sinusoidal model was used to generate a sample of 30 observations, except that the frequency was set at twenty cycles over the range of 30 sample pairs. With a frequency this high, successive X -points have their Y values far apart. The above procedure based on searching for maximal z was then run. This gave $K^* = 3$, with z^* , the maximal z , equal to 2.9263 and the proportion in 2000 values of $z^*(Y')$ which exceed z^* was .036, whereas the minimum z was -1.44 and the proportion of 2000 random minimum z values below -1.44 was 0.408. This is strong evidence against H_0 , and in favor of the hypothesis of a very complex relationship of some sort. In time series applications, negative autocorrelation would be an example of this phenomenon and the K -Optimal method could be used in this situation also.

3.3 Spaces of lists of symbols

Sample elements are sometimes described by lists of attributes, such as, for example, lists of keywords for documents, or personality checklists for persons. There are many other examples of data of this form, especially in the social sciences.

The following example illustrates how a test for association with such data may be carried out. The metric spaces M_1 and M_2 in this example will each be the space M of sub-sets of distinct elements from a finite set S . The distance d between two sets A and B will be just the number of elements by which the sets differ, or put another way, it is the number of elements in $AB' \cup A'B$. In particular and for simplicity and convenience, the sub-sets are just sub-sets of five distinct letters each taken from the first nine letters (the “attributes”) of the alphabet so that $S = \{A, B, C, D, E, F, G, H, I\}$. To illustrate if $X = H, K, C, A, I$ and $X' = C, B, H, K, J$ then $d(X, X') = 4$. Such elements from M will be called “words”.

A sample was prepared of 30 correlated pairs of such words, $(X_1, Y_1), (X_2, Y_2), \dots, (X_{30}, Y_{30})$, with each X_i being selected at random from M , with the associated Y_i being a random transformation of X_i produced as follows: First a fixed function g of S onto S , was applied to each of the letters in X_i to get five distinct letters, also an element in M . The result is a function, say g^* , of M onto M , defined by g . This function g^* might be thought of as the “true regression function”. To add random error to the regression, random mappings f' of S onto S were prepared by selecting at random two elements from the nine elements in S and interchanging them. Thus if the pair u, v was selected then $f'(u) = v$ and $f'(v) = u$, with $f'(u) = u$ otherwise. Such transformations can be generated independently of one another and applied as necessary to add “error” to the result from g .

For the illustration at hand, four such random mapping based on a random two letter interchange were applied to each of $g^*(X_i)$ to get Y_i .

The resulting sample of pairs is shown in Table 4. For convenience in perusal, the letters in the X words were alphabetized before applying the random transformation.

The reader is invited to make a judgment by inspection, as to whether or not there is some sort of association. It is perhaps not too difficult to tell that there is something there,

X	Y	X	Y
EFGHI	HEABG	BDEFG	ICBGE
CDEGH	BIFAD	ABCHI	IHC DG
ABCFG	IFDAC	ACDEG	HGBCA
ACGHI	GABDF	ABDGI	BHIGD
ACEFG	GFDCB	ABCFH	IFHED
BCEGI	ECFDG	ABEHI	GBCID
CDGHI	FBGAH	BFGHI	CEDAG
ABDFG	AHBEF	BCEFH	HFCE D
BDEFI	GBIHA	CEFHI	ADEGC
ABDFH	BDIEC	CEFGI	DHEGB
ABDGH	HABDI	BCFGH	AEDIF
ABCEF	EFCHA	ADEGI	EBCHG
BCEFI	CFGEH	BDEFG	HCBAD
ABCFI	IHFDE	BCDEG	HFACB
ACFGI	BFEHG	ABDFG	CHBEG

Table 4: Data for the List of Symbols

but the statistical task is to make some more precise and quantitative judgment.

The K -optimal test was applied and, with a significance probability of 0.0016, there is evidently quite strong evidence for association.

For a comparison, another test for association based on clustering was applied. This simple procedure suggested by MacQueen (1965,1967) consists of partitioning the X 's into K similarity groups and then measuring the Y variation in the partition of the Y 's induced by the X partition. Because the X 's in each group tend to be close to one another the associated Y values should also be relatively close within each set. Thus the Y within group variation defined by the sum of all pairwise distances within each set, summed over all sets, just as in the K -optimal test, may be used as a test statistic and evaluated by randomization, either by computer sampling or by a normal approximation.

This procedure was applied to the X 's in the data of Table 4, using a clustering procedure called KCENTERS, which is variant on the well known K -means procedure appropriate for metric data. The number of clusters was chosen to be $K = 7$ on the experiential basis that that this number is convenient for purposes of data perusal. The clustering procedure is described in Appendix 2.

The result after applying the within group variation tests to the associated Y clusters was a significance probability of 0.051 with 1,000 randomizations, substantially larger than from the K -optimal (with $K^* = 8$) significance probability of 0.0016, indicating that the K -optimal test is considerably more powerful.

The clustering method has the advantage that the clusters are available for perusal, and in fact the usefulness of the clustering as a method of data perusal is enhanced by the K -optimal test because it has reduced the risk that this analysis is not reading into the data something not really there.

3.4 Boolean regression

Consider variables x, y, r and s , where each is an expression or sentence which is either true or false. These can be combined in various ways using the basic connectives, “and”, “or”, “not” and “implies”, to form other expression and we let f be the resulting expression. For example, $f = [(x \Rightarrow y) \text{ OR } (r \Rightarrow s)]$. Interpreting 1 as “true” and 0 as “false” in the usual way, such functions f may be interpreted as true regression functions. Random “error” in observing f may be expressed by choosing a probability q and making $f = 1$ with probability q if f is true, and 0 with probability q if f is not true, and so $(1 - q)$ can be interpreted as the “noise” in the data set. The elements x, y, r and s are chosen to have truth values with certain probabilities.

With the 0/1 interpretation in mind any such f becomes a numerical valued function of the variables x, y, r, s themselves being assigned values 0 or 1 according to their truth or falsity.

To see if the K -optimal tests could detect such highly non-linear relationships, a number of samples were prepared with different functions, different values of q , and different sample sizes. The values of the predictor variables x, y, r, s were taken to be true or false with probability 0.5 and independent of one another. Only the results for $N = 100$, are given and these are in Table 5. The regression function used and the values of \hat{p} are shown in this table along with the significance probabilities provided by the K -optimal test.

Problem	k	N	K^*	Z^*	Noise	\hat{p}	Function
1bool1	4	100	9	-10.8090	0.10	0.0000	f_1
1bool2	4	100	8	-9.1486	0.15	0.0000	f_1
2bool1	4	100	8	-4.9724	0.10	0.0000	f_2
2bool2	4	100	6	-2.4402	0.15	0.1170	f_2
3bool1	4	100	30	-4.5658	0.10	0.0000	f_3
3bool2	4	100	12	-8.7540	0.15	0.0000	f_3

Table 5: Boolean Regression Functions

The three Boolean functions used were $f_1 = [(x \Rightarrow y) \text{ AND } (r \Rightarrow s)]$, $f_2 = [(x \Rightarrow y) \text{ OR } (r \Rightarrow s)]$ and $f_3 = [(x \text{ AND } y) \text{ OR NOT } (r \text{ AND } s)]$. There was no difficulty in detecting the presence of association with $q = .8$, even with samples of size 30, (not shown) except in the case of $f_2 = [(x \Rightarrow y) \text{ OR } (r \Rightarrow s)]$. The reason for this is that out of all the 16 possible patterns of values of x, y, r and s , only one is false. The value of f in this case becomes close to being independent of the other variables with a constant probability of 15/16 of being 1.

3.5 Alternative choices for K

As an alternative to using \hat{p}^* as a test statistic, we could use V itself, since small values of V also offer evidence against H_0 . This suggests that the value of K minimizing V might be a good value of K and the associated minimum value of V might be a good test statistic. This value will be called V^* . Significance probabilities could be calculated by brute force randomization just as with the K -optimal test.

This method was implemented and the results compared with the “ K - Optimal” test. In these comparisons the two methods performed somewhat similarly over a variety of simulated data sets. Table 6 shows the significance probabilities $\hat{p}_{V.\text{opt}}$ and the optimal K , labeled $K_{V.\text{opt}}^*$ for this method for some of the data sets used in Section 3 for comparison with a “rule of thumb” choice of K (discussed below). However, there is a tendency for the minimizing value of K to be small in some instances when with the K - Optimal test a large K is obtained. The reason for this appears to be that the larger values of K give a smaller variance for V

under randomization, with an attendant stronger significance probability. This explains why in those cases the K -Optimal method is superior to the V -Optimal method.

Problem	$K_{V.Opt}^*$	$\hat{p}_{V.Opt}$	K^*	\hat{p}	$\lceil N/4 \rceil$	$\lceil \sqrt{N} \rceil$	$\lceil \log_2 N \rceil$	$\Delta = 2$	$\Delta = 5$
lin0	16	0.3605	16	0.1155	0.1069	0.3004	0.3784	0.0865	0.1015
lin1	16	0.2025	16	0.0095	0.0160	0.0584	0.0946	0.0070	0.0110
s1n30std2	2	0.0035	26 ¹	0.0000	0.0004	0.0002	0.0001	0.0000	0.0000
s2n30std3	2	0.0130	26 ¹	0.0000	0.0010	0.0006	0.0003	0.0000	0.0005
sin25	5	0.0115	11	0.0000	0.0002	0.0003	0.0002	0.0005	0.0005
sin50	3	0.0395	11	0.0000	0.0107	0.0094	0.0069	0.0325	0.0150
1bool1	4	0.0175	9	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
1bool2	5	0.0135	8	0.0300	0.0000	0.0000	0.0000	0.0000	0.0000

Table 6: Results for Alternative Methods for choosing K .

This is illustrated by the data set sin25 where the optimizing K for V^* is 5 whereas K^* is 11. The relationship in this data set is quasi-linear and the larger value of K^* comes from the reduction in the variance of V . The advantage of smaller values of K comes from being able to respond to non-linear changes in the location of the mass of Y as a function of X , and this is not of great value in the linear cases.

The general issue here seems to be that choice of K is a decision under uncertainty. A given method for choosing K may be very good against a particular range of situations, and, when the data is in this range, a strong significance probability is obtained, but if the method is not powerful for a different range of situations where the data in question happens to lie, the method reports back in favor of the H_0 . The K -optimal test appears to be very unlikely to miss any relatively strong relationship because it is using an inappropriate value of K .

This issue may be clarified by comparing the K -optimal method against various “rules of thumb” which we previously found appealing. For example having K the smallest integer greater than or equal to \sqrt{N} has some appeal, as being among those rules which Stone (1977) found to give exact predictions as N becomes large. If a data set is one for which

¹These large values for K^* are not fully understood, but are occasionally observed especially with a regression function of a highly nonlinear nature.

this choice really works is at hand the significance probability is considerably enhanced. The data set lin5 (not shown in Table 6) illustrates this. With $N = 30$ the value of K^* was 6, and the rule of thumb with $\lceil\sqrt{N}\rceil = 6$ would have revealed a significance probability of 0.003210 whereas the K -optimal test gave a significance probability of only 0.0075. Two other rules of thumb that we thought to be of interest were $\lceil\log_2 N\rceil$ and $\lceil N/4\rceil$. For the “rule of thumb” experiments, we increased R to 100,000.

Each of these performed better than the K -optimal tests when the value of K^* was near these rule of thumb values but the K -optimal test outperformed these as soon as K^* was some distance from the rule of thumb value.

This leads to the possibility that there are other systematic search methods which are generally more powerful perhaps because they find good values of K but because they do less search and do not suffer as much from finding spuriously strong values of Z among the randomizations. For example, taking every third value of K might locate values of K which provided reasonably powerful tests but would suffer less from the search. With the final significance probability being maintained by the randomization perhaps even a uniform improvement over the K -optimal procedure would be achieved.

We have explored this possibility in a very small way. Taking every other value of K and every second value of K , every third value, up to every fifth value, did not give any improvement over the K -optimal test. These results for $\Delta = 2$ and $\Delta = 5$ (Δ represents the step size for the values of K considered) are displayed in Table 6. Of course, the case $\Delta = 1$ is just the K -Optimal procedure already in the table. The possibility of a fully sequential procedure, where the values of K are chosen on the basis of the results from earlier values is intriguing. Of course, as long as such a procedure is fixed in advance, the final significance probability will be maintained correctly on the basis of the “brute force” randomization.

But our feeling is that these other search procedures may lead into a morass of methodology which however correct it may be from a logical point of view, will gain little power over the K -optimal method and will lose considerable merit in terms of simplicity and intelligibility.

4 Conclusion

In the situations where the power of the K -Optimal test was compared to other tests on the basis of simulated samples, using significance probability as an index of power, the K -Optimal test was found to be at least as good or slightly superior, and in some instances distinctly superior in power. In the case of multiple regression superior power was evident for the F test, but this required that the multiple regression analysis be based on a correctly specified model. When the accuracy of the specification is unclear and its effect on significance probabilities produced by F is not known, the K -Optimal test may be well be preferred even in the multiple regression situation. It requires no specification and the meaning of the significance probability estimates is clear.

The power shown by the K -Optimal test is evidently due to the robustness of the Nearest Neighbor logic on which it is based. As was pointed out in the introduction, this logic assumes very little more than the principle of the continuity of nature, as asserted by the classical statement “like causes have like effects”. This basic idea, quantified using the abstract notion of distance, translates easily and directly into the assertion that if two X points are near, there is tendency for the associated Y point to be near. The nearest neighbor logic is just this and its fundamental nature cannot be overemphasized. Thus it is a truism that no two empirical situations are ever exactly alike except at the abstract level of mathematics, so similarity is the best we can hope for in using past to help predict the future and thus better cope with it. The K -Optimal test is a refinement of the nearest neighbor logic in that it provides a reasonable measure of association V and a way of putting its strength on the intrinsic scale called significance probability. The role of K in this is basically just that it expresses the amassing of evidence in some local way so that the relationship with Y is better revealed. It appears that the significance probability of the test for association is a direct measure of the extent to which the many individual facts represented by the X_i bear on the problem of predicting the Y_i , and this accounts for its value in choosing K .

That a very wide range of applied situations can be approached with the same logic is

a useful feature of the test. Learning to understand the randomization logic is its main cost. But the logic is exactly that of any non-parametric test, and these are widely used partly because of their ease of understanding. The only feature of the K -Optimal test which required some careful thought and practice before it becomes easy and clear, is that while the test statistic is based on a search over all possible values of a parameter of the method, K , doing this for all the randomizations keeps the significance probability as exact as desired. But once this logic is understood and accepted, the ease of application is very evident.

The most technical aspect of the method is the use of the normal approximation. That this would work well is intuitive to anyone who has studied the central limit theorem, which is found to be extremely robust in applications. But as was suggested above, the normal approximation, while helping to understanding why the test might be expected to work well, is not strictly speaking essential. The significance probabilities derive from the purely ordinal properties of the test statistic \hat{p}^* . Familiarity with the binomial distribution and unbiased estimation is about all that is required by way of mathematical training for use and interpretation of \hat{p} . That the test is based on the random pairings has an easy and straight forward interpretation. Pairing the Y 's randomly with the X 's is just a concrete simulation of what is meant by asserting they are independent. And when the simulation is used to deduce the conclusion that a totally unexpected consequence has been observed under H_0 , the subjective probability of this hypothesis drops accordingly, from common sense, from approximate application of Bayes theorem, and because this is the way the mind works. DeGroot (1973) has discussed the role of significance probabilities in the evaluation of empirical data, and shows that such subjective probability revisions taking account of significance probabilities are essentially correct subject to a few obvious provisions, even though they are not derived rigorously in the Bayesian tradition.

The K -Optimal test, in addition to being very general, is also basically simple and intuitive. Many data situations where some sort of test for association has been lacking, or might appear to require development of special and seemingly situation specific methods, can be treated directly by this test.

References

- Cover, T.M. and P.E. Hart (1967), “Nearest neighbor pattern classification”, *IEEE Trans. Information Theory*, Vol. IT-13, pp 21- 27, January, 1967.
- Cover, T.M. (1968), “Estimation by the nearest neighbor rule”, *IEEE Trans. Information Theory*, Vol.IT-14, No.1, January 1968
- DeGroot, M.H. (1973), “Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio,” *Journal of the American Statistical Association*, pp. 966–969.
- Diaconis, Persi (1988), “Group Representations in Probability and Statistics,” Institute of Mathematical Statistics, Hayward California.
- Fix, E. and J.L. Hodges, Jr. (1951), “Discriminatory analysis, non- parametric discrimination, consistency properties,” *USAF School of Aviation Medicine*, Randolph Field, Tex., Project 21-49-004, Rept. 4, Contract AF41(128)-31, February.
- (1952) “Discriminatory Analysis, small sample performance”, *USAF School of Aviation Medicine*, Randolph Field, Tex., Project 21-49- 004, Rept. 11, August.
- Forsythe, A.B., L. Engelman, R. Jennrich, (1973) “A Stopping Rule for Variable Selection in Multiple Regression,” *Journal of the American Statistical Association*, pp. 75–77.
- Good, Phillip (1994), “Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses”, Springer-Verlag, Berlin.
- MacQueen, J. (1967), “Some methods for classification and analysis of multivariate observations”, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, LeCam, L.L. and Neyman, J., Eds., 281-296.
- MacQueen, J. (1968), “Methods for statistical analysis and testing in practical metric spaces”, *Invited Paper*, Western Regional Meetings of the IMS, Missoula, Montana, June, 1967. Abstract, Institute of Mathematical Statistics Bulletin, Vol. 39, No. 2, p. 687, 1968
- MacQueen, J. (1991a), “An Approximate Randomization Test for Nearest Neighbor Regression with Metric Space Variables”, *Institute of Mathematical Statistics Bulletin*, Abstract

91t-44, p. 212.

MacQueen, J. (1991b), “An Unbiased Estimate of the Variance of the Sample Variation for Metric Space Data”, *Institute of Mathematical Statistics Bulletin*, Abstract 91t-46, p. 560.

Manly, Bryan F.J. (1991), “Randomization and Monte Carlo Methods in Biology”, Chapman and Hall, London.

Marriott, F.H.C. (1979), “Barnard’s Monte Carlo tests: how many simulations?”, *Applied Statistics* **27**, 75–77.

Stone, C.J. (1977), “Consistent nonparametric regression”, *Annals of Statistics* **5**, 595–620.

Sverdrup-Thygeson (1981), H., “Strong Law of Large numbers for Measures of Central Tendency and Dispersion of Random Variables in Compact Metric Spaces,” *Annals of Statistics* **9**, 141–145.

Appendix 1: Mean and Variance of V'

The first two moments of the distribution of V' can be readily calculated using the following quantities:

$$\begin{aligned}
 N_\ell &= N(N-1)\cdots(N-\ell) \\
 K_\ell &= K(K-1)\cdots(K-\ell) \\
 A &= \sum_{i=1}^n \sum_{i<j}^n d_Y(Y_i, Y_j) \\
 B &= \sum_{i=1}^n \sum_{i<j}^n d_Y^2(Y_i, Y_j) \\
 C &= \sum_{i=1}^n \left(\sum_{j=1}^n d_Y(Y_i, Y_j) \right)^2 \\
 D_1 &= 2B/N_1 \\
 D_2 &= (C - 2B)/N_2 \\
 D_3 &= 4(A^2 - C + B)/N_3 \\
 u_{ij} &= |S_i \cap S_j| \\
 H &= \sum_{i<j}^n u_{ij} \\
 H^2 &= \sum_{i<j}^n u_{ij}^2 \\
 C_1 &= NK_1/2 + H^2 - H \\
 C_2 &= NK_2 - 2(H^2 - H) \\
 C_3 &= NK_3/4 + N_1(K_1/2)^2 + H^2 - H - 2H(K-1)^2
 \end{aligned}$$

Then the first and second moment are given by the expression:

$$EV' = \frac{2A}{N_1}$$

and

$$EV'^2 = \frac{4}{K_1^2 N^2} (D_1 C_1 + D_2 C_2 + D_3 C_3)$$

The formulae are essentially from MacQueen (1991a), but the transcription error in (1991a) has been corrected here.

Appendix 2: The Method of KCENTERS

This method first partitions the X points into K clusters, and then does a version of analysis of variance appropriate to the corresponding K groups of Y values in M_2 . That is, let T_i be the set of Y values associated with the i th set in the partition of the X 's and let $W_i = \sum d_2(Y_i, Y_j)$ where the sum is taken over all pairs of distances between the elements in T_i . The test statistic $W = \sum W_i/r_i$ where r_i is the number of elements in the i th set, is then evaluated by a randomization test, H_0 being rejected if $W \leq W_\alpha$ where the probability of a random W falling below W_α is α (See MacQueen 1965).

The clusters were obtained by a method called “ K -Centers” which is a version of “ K -means” appropriate for metric space data. A center for a set of elements (in M_1 in this instance) is any point in the set which minimizes the sum of the distances from itself. The K -centers program starts with a random set of K distinct elements from the X 's, and with these as initial centers and finds a partition by sorting the remaining elements on the basis of nearness to these initial centers. The centers of these sets are then determined, and a new partition of the basis of nearness to these is found, etc. The total distance of the elements from their nearest center, usually called the “within-group variation” can only decrease, so the process converges in a finite number of iterations. But the final centers depend on the initial choice of centers, so commonly one tries a number of initial centers, and selects finally the best one using W variation as a criterion. There is a positive probability of finding an optimal set of centers as measured by W since one might select an optimal set at the beginning each iteration.