Res. Lett. Inf. Math. Sci. (2003) **4**, 1-18
Available online at http://iims.massey.ac.nz/research/letters/

A Versatile Heuristic Test for Independence

Paul J. Bracewell
IIMS, Massey University Albany Campus, Auckland, N.Z.
p.j.bracewell@massey.ac.nz

Abstract
A functionally independent, heuristic test of independence is developed, explored and validated. Initially, a parametric method is described, before superior non-parametric methods that account for variations in skewness, kurtosis and sample size are obtained. This evolves into a multivariate procedure, which enables the importance of input variables to be established.

Introduction
The Half Moon Statistic (HM) is a powerful and versatile tool for use in exploratory data analysis. Primarily, this new statistic is for the detection of non-independence between two variables. The HM measure, like correlation, increases monotonically as the strength of association between variables increases. What makes this statistic so useful is the fact that no model assumptions are necessary.

Furthermore, the usefulness of the HM statistic is not limited to measuring the association between pairs of variables. The HM statistic is expanded to cater for the multivariate case (MHM), by considering the overall impact a set of input variables has upon an output variable, allowing calculation of the significance of a multiple input-single output association.

Functional independence is the advantage that the HM statistic has over conventional methods for determining the strength of association between variables. That is, knowledge of the nature of the function describing the association between the input and output is not required, which is a change from traditional model formulation and related statistical inference which "assumes the existence of a 'true' model (p.19)" [4].

In order to construct a much needed flexible method that determines whether a pattern exists, without assuming an underlying model [5], an algorithmic approach is adopted [3]. This provides emphasis on the workability with empirical justification of the proposed method, as adopted in data mining and related disciplines [7], rather than a mathematical proof that the methodology works.

Covariance, Correlation and the Coefficient of Determination
Consider two independent standard normal populations; a scatter plot of simulated data will show a rough circle-shaped object or swarm consisting of individual points. Any significant functional relationship between these two variables *will* distort the circle. It is this disruption to the circle that we wish to measure with the HM statistic.

Typically, covariance, correlation and the coefficient of determination are methods used to quantify the strength of association between variables. These methods are related as discussed in most introductory statistical texts.

The covariance between two random variables (X,Y) is given by:

$$\mathrm{Cov}\,(X,Y) = E[(X - \mu_X)\,(Y - \mu_Y)]$$

A non-zero value of the covariance is evidence of non-independence [8]. The scaling of the X-Y variables by subtracting the respective means focuses the measure on the relative dispersal from the centre of the swarm. Whilst it is convenient to regard the covariance as an index of dependence, this value can be zero even when the variables in question are not independent [8] due to a non-linear relationship. Covariance is not scale invariant in that a linear transformation of the X and Y data will alter the covariance. Additionally, this measure is restricted to only two variables. However, this covariance can be standardised to produce a dimensionless measure (correlation).

Scale invariance is in the coefficient of linear correlation, which is a ratio obtained by dividing the covariance by the square root of the product of the variances [8], as shown below:

$$\mathrm{Corr}\,(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

This scales the measure such that it has meaningful bounds [-1,1]. "The correlation between variates is a measure of the departure from circularity of the projection of the swarm (after standardisation) on to the plane formed by the two variate coordinate axes. In other words, it is a measure of the eccentricity of the ellipse so formed" [9]. As with covariance, calculation of the correlation coefficient compares observations with the centre of the ellipse, or swarm. The measurement of the eccentricity of an ellipse restricts meaningful association measures to linear relationships only (see also [8]).

The coefficient of determination is a measure of dependency for linear and non-linear relationships, defined as follows;

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

where e_i is the ith residual, obtained by subtracting the fitted model value from the ith observed value y_i. The coefficient of determination can be interpreted as the proportion of the total variability that is explained by the model and in the case of two variables (X,Y) it is just the square of the correlation coefficient – hence the name 'r^2'. It can be used for any model form with any number of predictor variables, and is the most commonly used descriptive measure of the strength of a relationship [11]. However, to obtain the residual, the fitted values must first be obtained, requiring that the functional form of the model family be known.

If the model family of the functional form describing the association between two variables is unknown, then the previously described methods cannot be applied to achieve a measure of association. In the following sections it will be demonstrated that by shifting the focus from the centre of the swarm to the perimeter of the swarm a functionally independent measure of association can be derived.

Underlying Philosophy for the Half-Moon Statistic

The introduction of a functional relationship between two standard normal variables distorts the circular swarm projected on to the plane formed by the two variate co-ordinate axes. This distortion should be measurable, regardless of the functional form of that relationship. This distortion will cause a change in the variance associated with the distances between the points and the circle circumference.

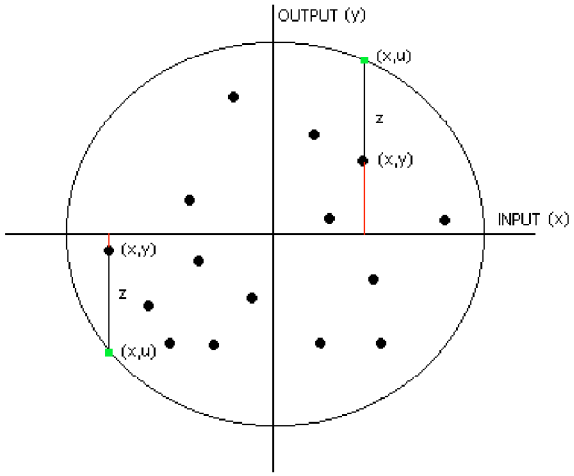

Figure 1: Geometrical Representation of the Half-Moon Statistic

Consider the vertical distance (Z) from a specified point (X,Y) to a point, (X,U), on an enveloping circle of predefined radius, r, as shown in figure 1. Then the variance of Z can be used to measure the strength of the X-Y relationship as the distribution of distances will differ depending on the strength of the association. It will be shown that the HM statistic is not sensitive to the value of r, for r sufficiently large ($r \geq 10$).

Mathematical Background for Bivariate Case

Figure 1 illustrates the geometrical derivation of the half-moon statistic for the bivariate case. For every (x,y) pair the magnitude of z is calculated by subtracting the absolute value of the output y from the absolute value of u, where u is the y co-ordinate for a point on the circle perimeter with input x. The half-moon statistic is defined using the sample standard deviation for z, calculated using all available (x,y) pairs. Given that the data has been standardised with a mean of zero and a standard deviation of one, under the initial imposition of independence, consider a circle centred at (0,0) with radius r, where r is chosen so that all the (x,y) points lie inside the circle. Then, if x_j is the jth observation of the input variable and y_j is the jth observation of the output variable,

$$x_j^2 + u_j^2 = r^2 \qquad (1)$$

$$u_j = \sqrt{r^2 - x_j^2} \qquad (2)$$

$$z_j = u_j - |y_j| \qquad (3)$$

with sample standard deviation:

$$s_Z = \sqrt{\frac{\sum_{j=1}^{j=k} z_j^2 - k\bar{z}^2}{k-1}} \qquad (4)$$

Then the half moon statistic (HM) is defined as:

$$HM = s_Z - \sqrt{Var(Z)} \qquad (5)$$

where Var(Z) is the population variance for Z, the vertical distances to the enveloping circle. The use of the absolute value for the output y has the effect of projecting the points into half a circle (the upper half); hence the name, half-moon statistic. The asymptotic zero-mean [10] feature of this statistic is crucial for creating a robust statistic for the non-parametric and multivariate extensions to follow.

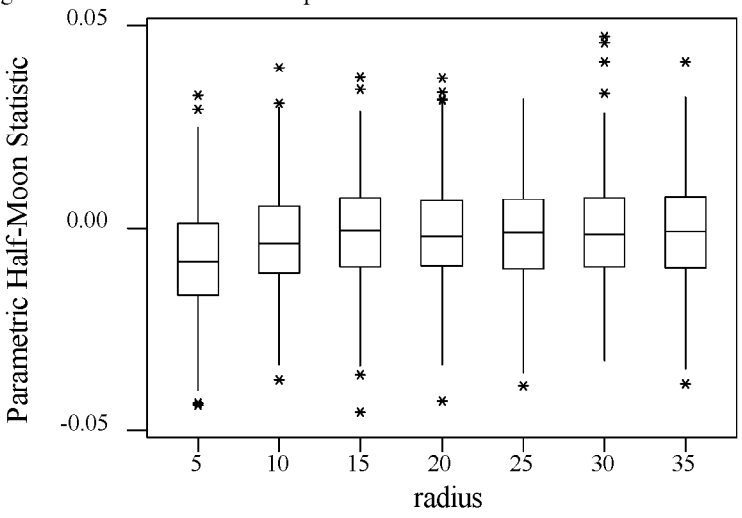

Figure 2: Boxplot of the Parametric Half-Moon Statistic for Varied Radii

The adoption of a particular radius for the univariate parametric statistic presented here is arbitrary, although it must exceed the magnitude of each input observation. This effect is illustrated in figure 2, with boxplots for seven different radii, calculated using 500 parametric half-moon statistics, each obtained from 500 observations from independent standard normal populations. The variances are equivalent, confirmed by a Levene's test for equal variances (p=0.609). A radius of 10 is adopted in this paper.

A parametric test for independence is created using the distributional properties of the HM statistic when the input-output variables are independent, which is examined before a more flexible (and suitable) non-parametric measure is developed.

Parametric Test for Bivariate Case

The half-moon statistic is defined using the standard deviation of Z. When X and Y have independent standard normal distributions the expected value for the HM statistic is easily calculated by using the two independent terms indicated in (3). The probability density function associated with $|Y|$ is a folded standard normal, defined as follows:

$$f_{|Y|}(y) = \frac{2\exp[\frac{-y^2}{2}]}{\sqrt{2\pi}}, \; y>0$$

with $E(|Y|)=2/\sqrt{2\pi}$ and $Var(|Y|)=1-2/\pi$ and the probability density function associated with X^2 is chi-squared with one degree of freedom.

The expected value of $\sqrt{r^2-X^2}$ is approximated using a Taylor Expansion [1]. It can be shown that for $X \sim N(0,1)$,

$$\begin{aligned} Var(\sqrt{r^2-X^2}) &= E(\sqrt{r^2-X^2})^2 - E^2(\sqrt{r^2-X^2}) \\ &= E(r^2-X^2) - E^2(r-\frac{X^2}{2r}-\frac{X^4}{8r^3}-...). \end{aligned}$$

Because the higher moments of the normal distribution become negligible for a normally distributed random variable, the variance becomes,

$$\begin{aligned} Var(\sqrt{r^2-X^2}) &\approx (r^2-1)-\left(r-\frac{1}{2r}-\frac{3}{8r^3}\right)^2 \\ &\approx \frac{1}{2r^2}-\frac{3}{8r^4}-\frac{9}{64r^6}, \end{aligned}$$

which provides a good approximation [13]. Therefore for two independent, standard normal (X,Y) populations the expected variance is,

$$Var(Z) \approx \left(1-\frac{2}{\pi}\right)+\left(\frac{1}{2r^2}-\frac{3}{8r^4}-\frac{9}{64r^6}\right). \qquad (6)$$

The parametric half-moon test statistic for testing independence between two standardised normally distributed variables is then defined as,

$$HM \approx s_Z - \sqrt{\left(1-\frac{2}{\pi}\right)+\left(\frac{1}{2r^2}-\frac{3}{8r^4}-\frac{9}{64r^6}\right)}.$$

The use of the expected variance, Var(Z), as a scalar produces a generic half-moon statistic with zero mean, as displayed in figure 3.

The rejection region for the hypothesis of independence lies in the tails of this distribution because input-output dependence increases the absolute value of the half-moon statistic.

A simulated distribution of 800 parametric HM statistics under the null hypothesis of independence, is shown in figure 3 for a sample size k of 500. The distribution of HM under the assumption of normality appears normally distributed - which is substantiated by the Anderson-Darling test for normality ($p=0.337$) – with a mean of zero, and a standard deviation of 0.0125. This enables appropriate critical values for the parametric test to be established. That is when X and Y are standard normal then for a 5% significance level and a sample size of 500, we must reject the null hypothesis of independence whenever HM lies outside the range [-0.0242, 0.0242] (analogous to a coefficient of determination of 0). The corresponding critical value for independence at the 5% level of significance for the coefficient of determination is 0.0077 ($\pm 0.0877^2$), obtained using Fisher's approximation [8]. The effect of k on the critical values for independence is examined later for the non-parametric case.

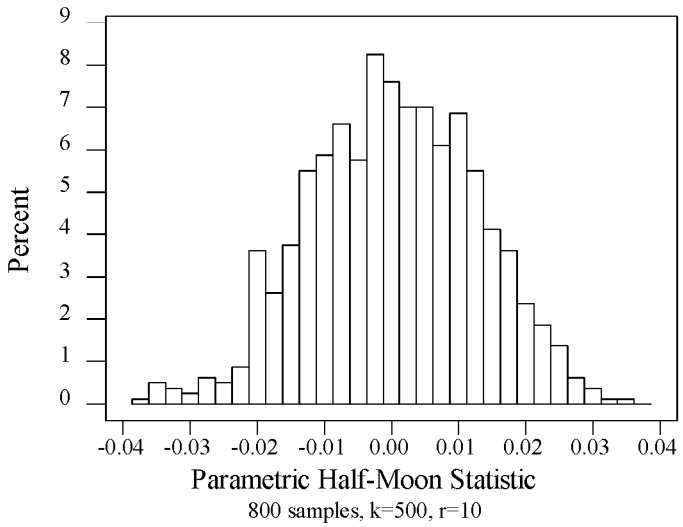

Figure 3: Histogram of Parametric HM Assumming Independence and Normality for X and Y

Whilst assessing the HM statistic as a test for independence, it is important to validate the measure using accepted measures of association. This is done for linear relationships of the form $y = b.x + c.e$, for $x \sim N(0,1)$, $e \sim N(0,1)$, $b^2 = 0.0, 0.1, \ldots, 1.0$ and $c^2 = (1-b^2)$. Plotting the parametric HM statistic against the coefficient of determination in figure 4 indicates a monotonically increasing relationship, supporting the use of this statistic as a measure for the strength of association between variables in the linear case, when X and Y are normally distributed. Critical values for the HM statistic are examined in the following sections.

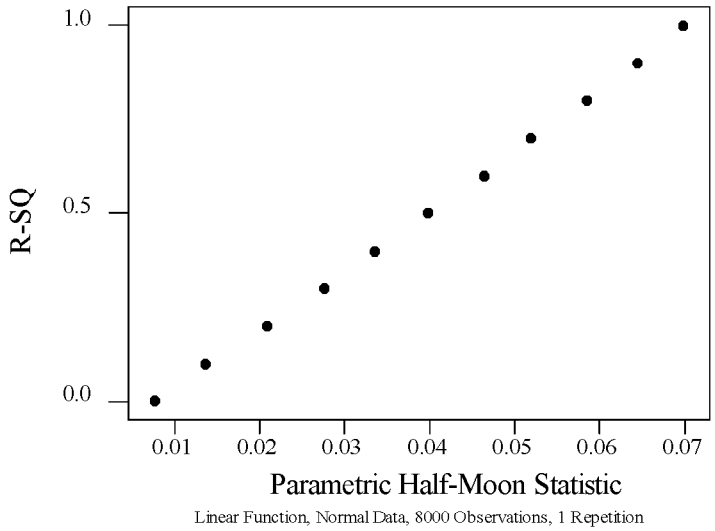

Figure 4. Plot of the Parametric Half-Moon Statistic against the Coefficient of Determination

The accuracy of the approximations introduced in this section to describe the properties of the parametric HM statistic are not dwelled upon, as the following sections will show that the non-parametric method is a superior test for independence.

Disadvantages of Parametric Test

Whist the previous sections have demonstrated the workability of the parametric half-moon statistic for linear relationships, the dependence on the normal distribution produces an inflexible method that is vulnerable to slight departures from normality and the impact of outliers. Additionally, a change from a linear to a non-linear relationship between X and Y is not explained well due to changes in the variation of the raw parametric statistic as a consequence of changes in the expected variance caused by alterations to the underlying distributions (equation 6).

The parametric test is vastly improved by removing the parametric properties from the calculation of the half-moon statistic. The non-parametric half moon statistic is a highly flexible method, which compensates for changes in the type of function determining the X-Y variable association by considering the unique characteristics of the distribution that result from the association.

Non Parametric Test for Bivariate Case

When the distributions of X and Y are unknown, a resampling approach is used to define a non-parametric test of independence for the half-moon statistic. The test statistic is defined as

$$HM = s_Z - \bar{s}_Z$$

where $\bar{s}_Z$ is the average of the standard deviations of Z calculated from m (non-parametric resampling coefficient) independent samples selected without replacement from the standardised Y data, with the standardised X data unchanged. The resampling of the X data is used to ensure that $\bar{s}_Z$ is calculated for independent X and Y samples. Figure 4 shows the distribution of this non-parametric HM statistic when X and Y are independent standard normal variables when m=3.

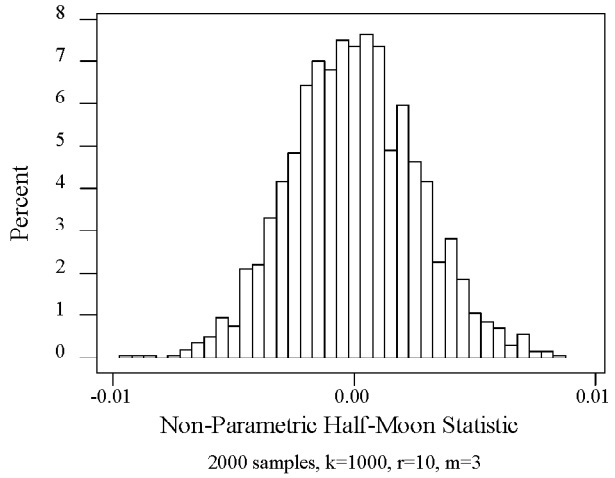

Figure 5: Histogram of Non-Parametric HM under Assumption of Independence and Normality

A simulated distribution of 2000 non-parametric HM statistics generated from independent X-Y variables, shown above in figure 5, seems normally distributed with mean, 0, and standard deviation, 0.0027 for a sample size of 1000 (p=0.178 for the Anderson-Darling normality test), suggesting critical values of ±0.0053 at the 5% level of significance. Due to the resampling approach adopted in the construction of the HM statistic, the standard deviation of this statistic is dependent on sample size.

Due to the sampling process, the choice of m impacts on the standard deviation of the non-parametric HM statistic, the type II error and also the time taken to compute the half moon statistic. That is, as m increases, the variance of the HM statistic decreases and consequently smaller critical values are obtained. Reducing the magnitude of the critical values decreases the likelihood of accepting a null hypothesis of independence when significant dependence is present (type II error). Increasing m requires more calculations to be performed, thereby increasing the computational time. This effect is minimal in the testing of actual data, but it is an issue when exploring the workability of the HM statistic through multiple simulations.

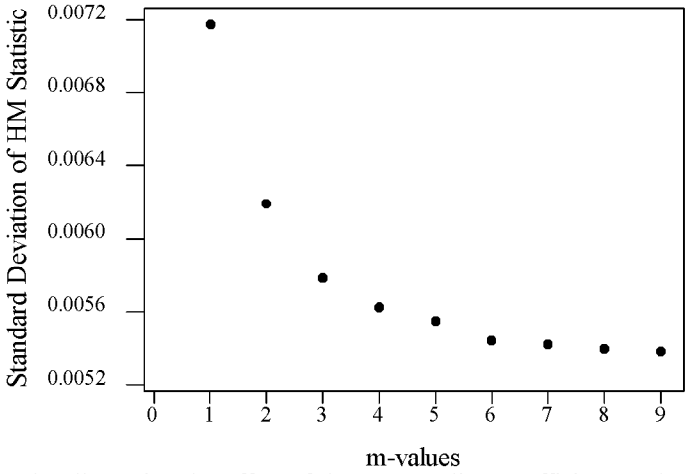

Figure 6. Plot Illustrating the Effect of the Re-Sampling Coefficient on the Sample Standard Deviation for the HM Statistic

However, as shown in figure 6 for 100 observations from standard normal data, the effect of reduced HM standard deviations as a result of increased m becomes less pronounced for $m \geq 3$. Figure 6 illustrates the effect of decreasing standard deviation for 200 half-moon statistics obtained from 100 observation, given an increase of the re-sampling coefficient, m. To reduce the computational time and minimise the type II error of the half moon statistic, the re-sampling coefficient is set at 3 for this paper. Altering of m will alter the critical values examined in this paper.

The non-parametric test provides a more precise test for independence than the parametric test because the idiosyncrasies of the data are retained. That is the unique characteristics of the data are incorporated into the calculation of HM, reducing the variability of HM, thereby reducing the probability of Type I and Type II errors. Comparing directly the parametric and non-parametric versions of the half moon statistic using the same data (standard normal data, k = 500, m = 3) highlights this effect. Whilst in both cases the mean is zero from 800 examples, the standard deviations of the methods are hugely different. The parametric test has a standard deviation of 0.01253 which is almost 3.5 times larger than the standard deviation for the non-parametric test, 0.00361. It is for this reason that the parametric test is not explored further.

With a mean, 0, and a standard deviation, ψ, critical values are calculated from standard normal tables at the required significance level. The use of normally distributed simulated data in figure 5 means that these critical values are only appropriate for variables that are approximately normal.

Effect of Sample Size

Due to the sampling process, the variability of the HM statistic is directly related to the sample size. Using 200 N(0,1) independent variables, repeated using 25 levels of varied sample size (50,100,…, 1250), provided the data for figure 7 and the subsequent regression analysis. Transforming sample size using the power (-1/2) produced the linear relationship evident in the figure below.

As a consequence of this linear relationship, appropriate critical values can be obtained with regard to the sample size. More importantly, because the distribution of the half-moon statistic is approximately $N(0,\psi^2)$, where ψ is determined by the sample size, p-values can be evaluated. From the regression analysis using the data from the above figure;

$$\psi = 0.0818174k^{-1/2},$$

where k is the sample size (p=0.000, R-SQ = 99.2). Therefore, for k = 8000, the following critical values apply (10%, ±0.0015; 5%, ±0.0018, 1% ±0.0024). Constructing a two-tailed test for independence at the 5% level of significance using 5000 independent N(0,1) X-Y pairings, with variable sample size (k = 50, 100,…, 1250) returned a misclassification error rate of 4.92% (246/5000).

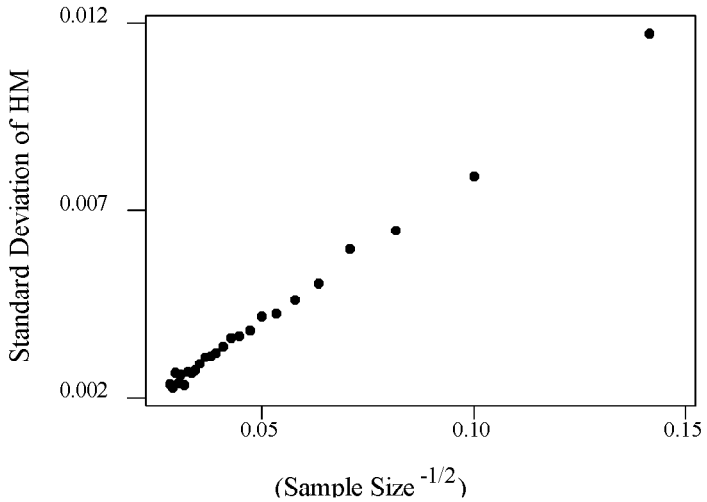

Figure 7: Relationship between Sample Size and Half Moon Variability

For the HM statistic to be effective, it needs to withstand skewness and kurtosis in the input variables, due to the influence of the encompassing circle. The effect of skewness and kurtosis is established empirically in the following section.

Effect of Skewness and Kurtosis

Preliminary analysis revealed that positively skewed data tended to increase the standard deviation of the HM statistic. The use of several different types of distributions and altering the parameters of these distributions allowed the impact of skewness and kurtosis upon the HM Statistic to be examined. As illustrated in figures 8 and 9, there is a relationship between the coefficients of skewness and kurtosis, and the HM statistic. The sample coefficients of skewness, ς_3, and kurtosis, ς_4, are defined as follows [12],

$$\varsigma_3 = \frac{g_3}{\left(\sqrt{g_2}\right)^3},$$

$$\varsigma_4 = \frac{g_4}{\left(\sqrt{g_2}\right)^4},$$

where g represents the qth sample moment of x about the sample mean for k observations as follows,

$$g_q = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^q}{k},$$

such that normally distributed data has a skewness coefficient of 0 and a kurtosis coefficient of 3.

The coefficient of kurtosis was transformed by first taking the natural log and then the square root, to produce figure 9. When the kurtosis of the input distribution exceeds approximately 40 (transformed kurtosis ≈ 2) the linear relationship that exists between the transformed kurtosis and the square root of the standard deviation of the HM statistic deteriorates. Consequently the corrected statistic is not suitable for data with high kurtosis.

The linear relationship evident in figures 8 and 9 allowed a correction factor for independent X-Y distributions. Figure 10 shows the combined effect of kurtosis and skewness upon the standard deviation of the HM statistic from a three-dimensional perspective. The *D*istribution corrected *HM* statistic (DHM) was obtained via regression analysis using the 34 observations detailed table 4 (p=0.000, R-SQ = 99.0). Distributions with high skewness and kurtosis apparent in the previous figures were omitted from the regression analysis which are italicised in table 4.

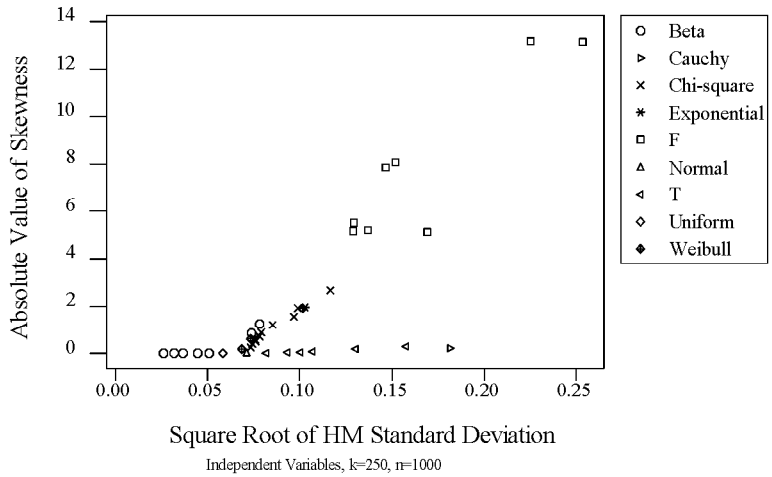

Figure 8: Relationship Between the HM Statistic and Skewness

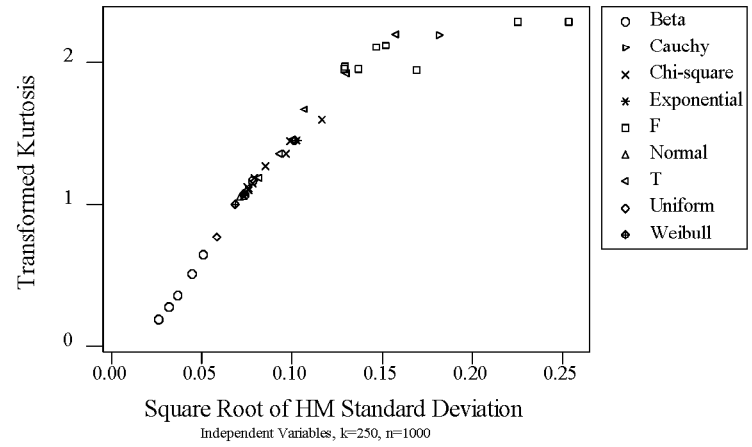

Figure 9: Relationship Between the HM Statistic and Kurtosis

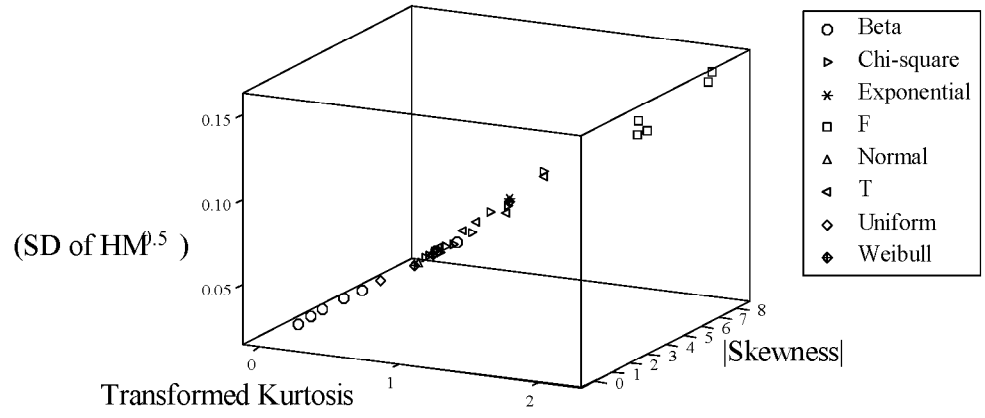

Figure 10: 3D Plot Demonstrating Relationship Between HM, Skewness and Kurtosis

Distribution	Description	Mean(Skewness)	Mean(Kurtosis)	SD(HM)
Beta	B(0.01,0.01)	0.0042	1.0344	0.0007
Beta	B(0.05,0.05)	0.0090	1.0780	0.0010
Beta	B(0.1,0.1)	0.0017	1.1352	0.0013
Beta	B(0.25,0.25)	-0.0059	1.2971	0.0020
Beta	B(0.5,0.5)	0.0013	1.5127	0.0026
Beta	B(2,0.5)	-1.2468	3.8374	0.0061
Beta	B(3,1)	-0.8559	3.1009	0.0055
Chi-square	Chi(60)	0.2541	3.0728	0.0054
Chi-square	Chi(30)	0.3664	3.1894	0.0055
Chi-square	Chi(25)	0.4998	3.3446	0.0057
Chi-square	Chi(120)	0.5482	3.4168	0.0058
Chi-square	Chi(20)	0.6028	3.5018	0.0057
Chi-square	Chi(15)	0.7180	3.7151	0.0061
Chi-square	Chi(10)	0.8650	4.0653	0.0063
Chi-square	Chi(5)	1.2131	5.0528	0.0073
Chi-square	Chi(3)	1.5531	6.3498	0.0093
Chi-square	Chi(2)	1.8941	8.0061	0.0098
Chi-square	Chi(1)	2.6573	12.8224	0.0136
Exponential	e(1)	1.9221	8.2008	0.0106
F	F(1,3)	8.0492	89.3096	0.0231
F	F(1,5)	5.5294	48.6389	0.0167
F	F(3,3)	7.8587	86.5426	0.0216
F	F(3,5)	5.1898	45.9162	0.0187
F	F(5,5)	5.1412	45.4764	0.0166
Normal	N(0,1)	0.0033	2.9960	0.0050
T	T(2)	0.1755	40.5614	0.0170
T	T(3)	-0.0574	16.1598	0.0114
T	T(4)	-0.0321	8.3991	0.0100
T	T(5)	-0.0375	6.3354	0.0087
T	T(9)	0.0090	4.0830	0.0067
Uniform	U[0,1]	0.0058	1.8117	0.0034
Weibull	W(1,2)	1.9182	8.2040	0.0103
Weibull	W(2,1)	0.6101	3.1703	0.0054
Weibull	W(3,0.5)	0.1611	2.7139	0.0047
Cauchy	*C(0,1)*	*-0.2239*	*125.3620*	*0.0329*
F	*F(1,1)*	*13.1916*	*188.6945*	*0.0508*
F	*F(3,1)*	*13.1447*	*187.2275*	*0.0642*
F	*F(8,3)*	*5.1173*	*44.8449*	*0.0286*
T	*T(1)*	*-0.2680*	*126.1929*	*0.0248*

Table 4: Skewness and Kurtosis Data

As a result of the regression analysis, the correction for the HM statistic in the presence of skewness and kurtosis is provided below:

$$DHM = \frac{HM}{\left(0.013391 + 0.0016113 \mid \varsigma_3 \mid +0.057329\sqrt{\ln(\varsigma_4)}\right)^2},$$

where ς_3 is the coefficient of skewness and ς_4 is the coefficient of kurtosis. Similar to the HM statistic, the DHM statistic is affected by sample size. Consequently, this more robust statistic is distributed approximately $N(0, \kappa^2)$ for independent X-Y pairings, where $\kappa = 15.2752k^{-1/2}$ and k is the sample size. As before, the formula for κ was obtained using simulated normal data for 25 sample sizes (p=0.000, R-SQ = 99.2). This information is then used to calculate p-values, indicating the likelihood that the association between X and Y variables differ from the null hypothesis of independence. Importantly, the HM and DHM statistics are equivalent for normally distributed data ($\varsigma_3 = 0$, $\varsigma_4 = 3$), after correction for sample size. That is both become standard normal distributions with means of zero and equivalent standard deviations following the adjustments for sample size.

Using the DHM statistic to create a two-tailed test for independence at the 5% level of significance using 5000 independent N(0,1) X-Y pairings, with variable sample size (k = 50, 100,…, 1250) returned a misclassification error rate of 4.76% (238/5000). Using the skewed data from figures 8 and 9 produced an error rate of 5.21% (2033/39000). These errors are obviously very close to the required 5% level of significance.

Examining Non-Linear Functions

The strength of the HM statistic is functional independence for (X,Y) variable pairs. That is, measures that are reflective of association are provided, regardless of the nature of that relationship. To illustrate this effect, coefficients of determination are simulated for three polynomial functions (linear, quadratic and cubic). These coefficients of determination are plotted against the corresponding HM statistics in the figure below. For example to obtain a coefficient of determination of 0.7 (70%), the following functions were employed, using 8000 observations from standard normal data for x and e;

Linear: $y = \sqrt{7}.x + \sqrt{3}.e,$
Quadratic: $y = \sqrt{7}.x^2 + \sqrt{6}.e,$
Cubic: $y = \sqrt{7}.x^3 + \sqrt{48}.e,$

where x is the input variable, e is noise and y the output of interest. Y has to be standardised before the HM statistic is calculated. Figure 11 shows clearly that there is a direct relationship between the appropriate coefficient of determination (R-Sq) and the obtained HM statistic, justifying the use of the HM statistic to measure the strength of association between variables, for polynomial functions. Importantly, the HM statistic increases monotonically as the coefficient of determination increases for each type of function, suggesting that the HM statistic is probably suitable for use as a test for independence when the input-output function is unknown.

In figure 11, The 95% confidence limits for independence for the coefficient of determination (0.0005) were established using Fisher's approximation [8]. The bounds for the half moon statistic (0.0018) were obtained previously.

The Type II errors of the HM and DHM statistics – the probability of failing to reject the null hypothesis of independence when the null hypothesis is false – is displayed in figure 12 for both one-tailed (upper-tail) and two-tailed tests.

200 HM and DHM statistics obtained from the X-Y association determined by three polynomial functions (linear, quadratic and cubic) with the coefficient of determination set at 50%, as described earlier in this section, produced 600 observations for each level of sample size (10, 25, 50, 75, 100, 150), using standard normal data. The type II error was established given a 5% level of significance. Clearly, the type II error decreases as the sample size increases. Consequently, both the HM stand DHM statistics are powerful given large sample sizes.

The univariate non-parametric half moon statistic is a robust measure of association. As a test statistic it has good coverage (accurate probabilities for type I error) and high power. The next section expands the univariate statistic to a multivariate case to test for independence in a system of inputs with respect to an output variable.

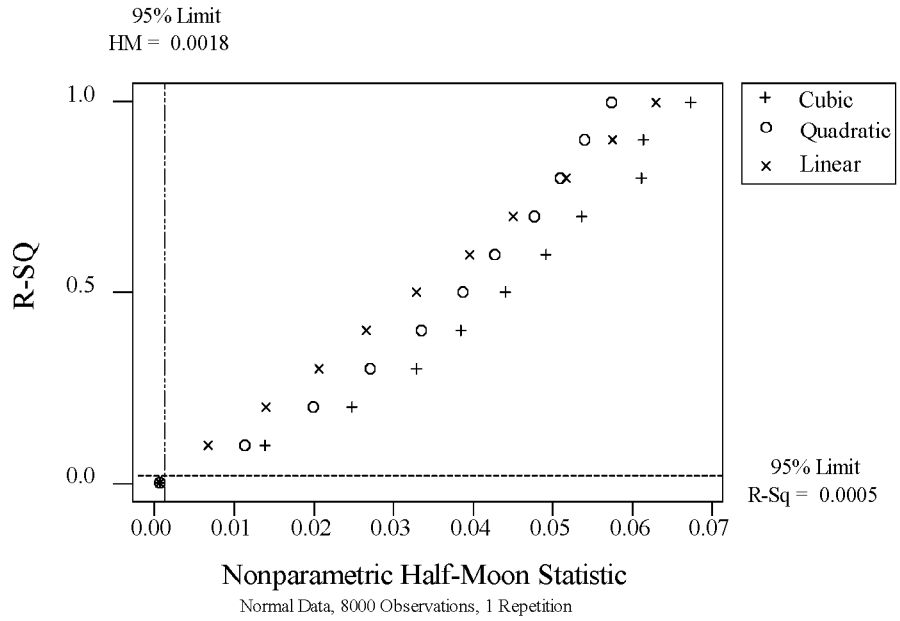

Figure 11: HM Statistic versus R-Sq for Polynomial Functions

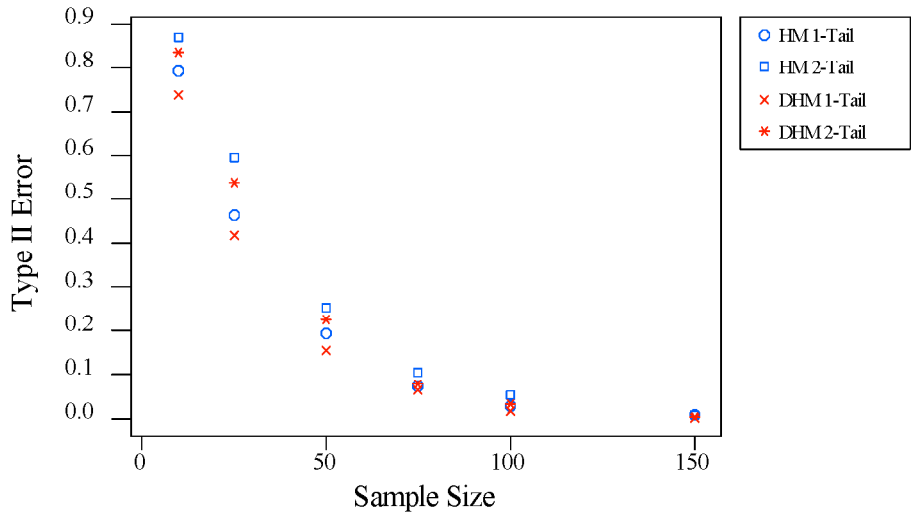

Figure 12. Type II Errors for Non-Parametric HM Statistic by Varied Sample Size

Expansion to the Multivariate Case (Multiple Inputs)

The half-moon statistic is easily extended to the multivariate case where instead of testing for independence between one input and one output variable, several input variables are considered with respect to an output. Let *n* be the number of input variables, then equation (2) is replaced with (7) to enable calculation of the multivariate half-moon statistic (MHM).

$$u_j = \sqrt{r^2 - \sum_{i=1}^{n} x_{ij}^2} \qquad (7)$$

for the jth observation and, as before, $z_j = u_j - |y_j|$ and $MHM = s_z - \bar{s}_z$.

Geometrically, instead of comparing points relative to the surface of a circle, comparison is relative to the surface of spheroid-type structures with the relevant dimensionality of the spheroid determined by the number of input variables to be considered. Because distance provides the basis for comparison (as shown in figure 1) increased dimensionality has no effect and the non-parametric MHM will still be one-dimensional, defined as in (5). However, increased dimensionality will affect Var(Z) unless r is adjusted.

Expansion to the multivariate case, assuming independence and identically distributed normal inputs (X_i) yields the following

$$\mathrm{Var}(u_j) = \mathrm{Var}\left(\sqrt{r^2 - \sum_{i=1}^{n} X_i^2}\right) = \mathrm{E}\left(\sqrt{r^2 - \sum_{i=1}^{n} X_i^2}\right)^2 - \mathrm{E}^2\left(\sqrt{r^2 - \sum_{i=1}^{n} X_i^2}\right).$$

Assuming the inputs are independent then the variance becomes

$$\mathrm{Var}(u_j) = \mathrm{E}(r^2 - X^2) - \mathrm{E}^2\left(r - \frac{X^2}{2r} - \frac{X^4}{8r^3} - \ldots\right)$$

where $X^2 \sim \chi_n^2$ so $E(X^2) = n$, and $E(X^4) = (n^2 + 2n)$. Given that r is sufficiently large, the higher moments become negligible. Taylor's expansion gives the following approximation for $\mathrm{Var}(u_j)$,

$$\mathrm{Var}(u_j) \approx (r^2 - n) - \left(r - \frac{n}{2r} - \frac{n^2(n+2)^2}{8r^3}\right)^2$$

$$\approx \frac{n}{2r^2} - \frac{n^2(n+2)}{8r^4} - \frac{n^2(n+2)^2}{64r^6}.$$

Such that $\mathrm{Var}(Z) = \mathrm{Var}(u_j) + \mathrm{Var}(|Y|)$. Therefore, for the multivariate case given independence and normality,

$$\mathrm{Var}(Z) \approx \left(\frac{n}{2r^2} - \frac{n^2(n+2)}{8r^4} - \frac{(n^2+2n)^2}{64r^6}\right) + \left(1 - \frac{2}{\pi}\right). \quad (8)$$

With increased dimensionality, the radius must be correctly defined such that equivalent values for HM and MHM statistics are obtained. This means that the radius r must be determined by the number of input variables, n, so that satisfying equations (6) and (8) provide equal variances for the HM and MHM statistics, thereby producing comparable HM statistics regardless of the dimensionality. This has important implications in the assessment of the interactions between input variables and combinations of input variables with an output variable. In particular, it means that the same critical values can be used for HM and MHM statistics, because Var(HM) will also be the same regardless of n.

For an initial (n=1) radius of 10, table 1 lists the comparable radii for more than one input variable. An approximation for the required radius is $\sqrt{10^2 n}$ when there are n inputs.

n	*Radius*	*n*	*Radius*
1	10.000000	8	28.026990
2	14.124152	9	29.687046
3	17.276279	10	31.250366
4	19.923110	15	38.006570
5	22.245653	20	43.568617
6	24.336842	25	48.340410
7	26.251940	30	52.530719

Table 1: MHM comparable radii

The effect of the comparable radii is illustrated in figure 13 using boxplots representing Z distances for various numbers of n simulated input variables given normality and independence and a sample size of 1000.

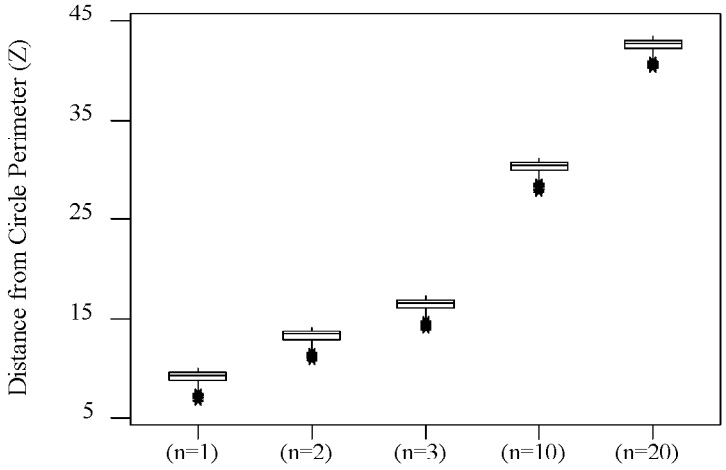

Figure 13. Boxplot Illustrating Unscaled z Distances for n=1,2,3,10 and 20

Clearly, as the radius increases in accordance with an expansion in the number of input variables (see table 1) the central location of the corresponding distribution increases. However, the central locality of the mean is irrelevant for the half-moon statistic which is defined in terms of Var(Z). Figure 14 displays the effect of re-scaling the measures by removing the expected value of Z, as defined earlier.

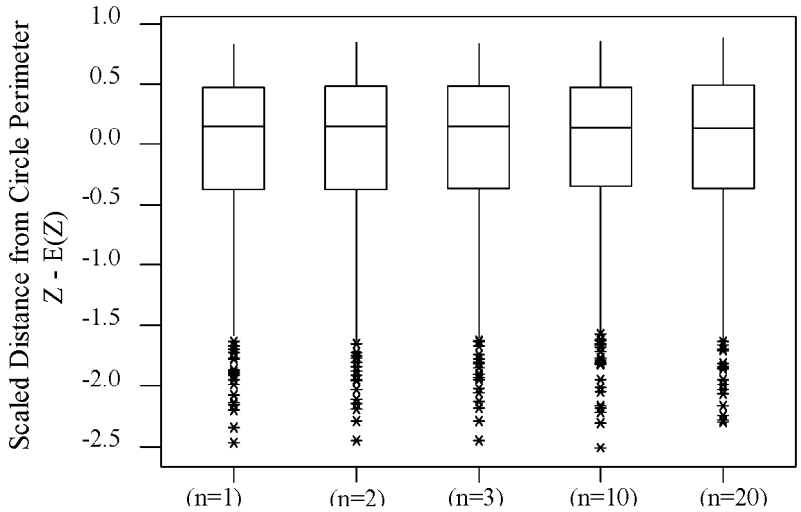

Figure 14. Boxplot Illustrating Scaled z Components for n=1,2,3,10 and 20

Figure 14 shows that Z is distributed equivalently regardless of the number of input variables provided the correct radius is adopted. The graphical evidence illustrating the similarity of the Z components between the varied input numbers is reinforced by a Levene's test for homogeneity of variance, indicating that there is no significant difference between the variances for Z of the different numbers of input variables (p=0.984). Figure 14 highlights the equivalence of variance, whilst figure 13 shows that the dependency on the radius relates to the expected value of Z and not the variance of Z. This is true regardless of the initial (n=1) value of r, and it highlights the robustness of the HM statistic to the r-value chosen.

When the distributions for the variables are unknown, the non-parametric procedure can be applied to the MHM statistic. However, further work is required to remedy skewness and kurtosis in the MHM procedure.

Importantly, as the distribution of the MHM statistic is equivalent to that of the HM statistic, the same approximate distributional properties hold ($N(0,\psi^2)$), enabling tests of independence of the system to be conducted. Furthermore, the MHM uses the same adjustments for sample size as the HM, as well as the same critical values.

Implementation of MHM to Calculate Relative Influence

Obtaining the equivalent variance, Var(Z), for variable numbers of input variables is crucial, as this allows combinations and interactions of variables to be assessed by removing variables and adjusting the radius accordingly. This is achieved by adopting comparable radii enabling the influence of input parameters or combinations of input parameters to be evaluated with reference to given standards regardless of the specified dimensionality.

This section will construct a relative influence measure for each variable, which provides two powerful results. Firstly, the use of the relative influence measure promotes parameter parsimony in statistical modelling. Secondly, as the combined effect of input variables can be assessed, combinations of variables can be chosen to maximise the information explained, whilst minimising the number of variables required. Related to parameter parsimony is the idea of variable redundancy. Examining the combined relative influence for the set of input variables allows the redundancy of that system to be established. Identifying redundancies is another exciting area for development, which is currently being investigated and showing great promise.

In the previous section it was detailed how the holistic effect of the input variables on an output could be established by calculating the overall MHM. More importantly, by omitting each variable one at a time, the relative influence an input variable has on the output variable, with respect to the other variables, can be established. This is an important step in establishing the redundancy of the input-output system. Further, interactions of input variables can be assessed by omitting specific combinations of input variables. The calculations of the core components defining the MHM are outlined in table 2. The nature of the formula denoting u_j in the defining calculation column of the table below is specified for computational ease. Once u_j is computed, calculation of the MHM proceeds as usual with $z_j = u_j - |y_j|$ and $MHM = s_z - \bar{s}_z$. Note MHM_{x1} = MHM with x1 omitted from the 'model' for Y and MHM is the MHM when all input variables are included in the model for Y. note that the 'model' is never specified because it does not need to be specified.

Variables of Interest	n for relative radii	Defining Calculation for u_j	Relative Influence Measure
ALL	n	$\sqrt{r^2 - \sum_{i=1}^{n} x_{ij}^2}$	-
x1	n-1	$\sqrt{r^2 + x_1^2 - \sum_{i=1}^{n} x_{ij}^2}$	$\frac{(MHM - MHM_{x1})}{MHM}$
x2	n-1	$\sqrt{r^2 + x_2^2 - \sum_{i=1}^{n} x_{ij}^2}$	$\frac{(MHM - MHM_{x2})}{MHM}$
Combinations of x1 & x2	n-2	$\sqrt{r^2 + (x_1^2 + x_2^2) - \sum_{i=1}^{n} x_{ij}^2}$	$\frac{(MHM - MHM_{x1 \times x2})}{MHM}$
Combinations of x1, x2 & x3	n-3	$\sqrt{r^2 + (x_1^2 + x_2^2 + x_3^2) - \sum_{i=1}^{n} x_{ij}^2}$	$\frac{(MHM - MHM_{x1 \times x2 \times x3})}{MHM}$

Table 2: Calculation of Key MHM Components using a Stepwise Procedure

The overall MHM provides an indication of the strength of the relationship between the input variables and the output. Given that MHM exceeds a suitable critical value indicates that the collective influence of the input variables is responsible for influencing the output variable. By omitting one or more input variables, those variables, or combination of variables responsible for the disruption can be identified. This procedure is similar to the stepwise procedures used in stepwise regression and best subsets regression.

As an illustration of the above table, given four input variables and an initial radius of ten, r is 19.923100 (see table 1) for the overall MHM. Calculations proceed as before in determining the non-parametric half-moon statistic. The relative influence explained by a variable is obtained by scaling the omitted MHM by simply subtracting the omitted MHM from the overall MHM and dividing through by the overall MHM as demonstrated in table 2. This measure can be used to rank the input variables in order of importance.

By omitting variables one at a time the impact of the chosen variables can be measured. The bigger the influence of an omitted input, the further the omitted MHM will be from the MHM calculated using all input variables, and the higher will be the relative influence measure. Conversely, the relative influence measure of unimportant variables will tend towards zero.

It is possible for these values to be negative, due to the sampling approach and the relationship between the one omitted MHM and the overall MHM. Negative relative influence measures indicate obvious redundancy. This has important implications in the evaluation of redundancies.

This information for the omitted MHM can be coupled with the HM and DHM statistics to gain an understanding of the model properties. The important aspect to arise from the calculation of the relative influence, is the indication of system redundancy, currently being developed.

Application

To demonstrate the effectiveness of these new tools in an applicable environment, an exploration of the commercially developed Eagle Rating system for individual rugby players is analysed. Eagle Sports (www.eaglesports.co.nz) collects a large amount of rugby data on a match by match basis from which a single match rating that summarises each individual's performance is computed. This single match rating is comprised of key performance indicators, or latent factors (KPIs) [2]. In this example the attack KPI for a cluster of individuals (midfield backs) is assessed with respect to 17 key input summary variables as indicated in table 3.

Variable	*HM*	*HM(p)*	*DHM*	*DHM(p)*	*RI*	*RI Rank*	*r*	*Skewness*	*Kurtosis*
Defence Beaten	0.0324	**0.00**	4.71	**0.00**	0.05	5	**0.63**	1.2	4.0
Errors	0.0235	**0.00**	3.35	**0.00**	0.03		0.21	1.1	4.2
Breaks	0.0933	**0.00**	8.63	**0.00**	0.21	1	**0.86**	1.8	10.4
Kicks	0.0013	0.79	0.12	0.90	-0.03		0.11	2.1	11.1
Running Metres	0.0587	**0.00**	8.44	**0.00**	0.12	3	**0.88**	1.0	4.2
Kicking Metres	-0.0087	0.08	-0.68	0.46	-0.05		0.04	2.8	15.9
Kicks Caught	0.0251	**0.00**	2.45	**0.01**	0.03		0.17	2.3	8.6
Breakdown Impact	0.0125	0.01	1.62	0.08	0.00		0.26	1.3	4.9
Passes	0.0510	**0.00**	5.52	**0.00**	0.09	4	**0.61**	1.5	7.1
Laybacks	0.0727	**0.00**	6.29	**0.00**	0.16	2	**0.76**	2.2	12.2
Tackles	-0.0035	0.47	-0.61	0.50	-0.04		-0.03	0.7	3.2
Missed Tackles	0.0071	0.15	0.91	0.32	-0.01		0.04	1.2	5.1
Tackle Assists	0.0033	0.50	0.37	0.69	-0.02		-0.02	1.5	6.6
Tries	-0.0011	0.82	-0.11	0.90	-0.03		0.11	2.3	7.4
Loose Ball Gained	-0.0047	0.34	-0.58	0.53	-0.04		0.16	1.4	5.4
Harassment	0.0108	**0.03**	0.87	0.34	-0.01		0.02	3.3	13.3
Infringement	0.0141	**0.00**	1.40	0.12	0.01		0.10	2.4	8.1

Table 3: HM Procedure Output for Attack KPI by Summarised Key Inputs

The data contains 279 observations relating to performances by starting midfield backs (second five eighths and centres) from the Super 12 competition, 2000.

Firstly, the univariate HM statistics can be used to interpret the output in the same way that loadings are used to interpret the factors in factor analysis [6]. Typically, if the absolute value of a loading (r) exceeds 0.5, then the variable is deemed to be having a significant influence upon that factor. In this instance, if a variable and the output are not independent as determined by the HM statistics, then that variable and the output are significantly associated.

As shown in table 3, the data is skewed and has moderate values of kurtosis. Consequently, the HM statistic is unsuitable for use with this data and the DHM must be used instead. The p-values of DHM indicate that *Defence Beaten*, *Errors*, *Breaks*, *Running Metres*, *Kicks Caught*, *Passes* and *Laybacks* are the variables that are significantly associated with the output (Attack KPI). As expected, these variables are predominately attacking based variables, which is congruent with the previous interpretation of this KPI. The use of the RI measure allows the importance of the variables to be ranked. The most important variable is the number of *Breaks* (1) made in each game, followed by the number of *Laybacks* (2).

The multivariate aspects of the HM procedure provide a holistic understanding of the inputs-output system. Importantly, there is significant association between this set of inputs and the output (MHM = 0.09, p-value = 0.00). That is the input-output relationship is not independent. This was expected given that the attack KPI is created by the combination of the significant variables mentioned previously.
Thus, this application has shown the wealth of information generated by the HM procedure, and shown how these statistics may be used.

Conclusion

This paper has introduced important tools for exploratory data analysis. Measurements for the degree of association between input-output variables (HM, DHM, MHM & RI) were developed and explored. Due to the simple nature of the mathematics involved, a MINITAB macro was developed to automate the implementation of the entire HM procedure.

The non-parametric half moon test statistic (HM) performs as well as the coefficient of determination in measuring the strength of a polynomial relationship between two variables, with the added bonus that a functional form for the input-output relationship does not have to be assumed. As a test for independence using standard normal input data, the statistic is approximately normally distributed (given the null hypothesis of independence) with mean, 0, and a standard deviation, ψ, that is determined by the sample size. Whilst the statistic does not determine the form of the model, it does indicate which input variables are associated with the output variable. This statistic was then modified to handle moderate skewness and kurtosis in the data (DHM). This statistic, assuming independence, is also normally distributed with a mean, 0, and a standard deviation, κ, which is dependent on sample size.

The univariate HM was then expanded to consider multiple inputs (MHM), enabling the interactions of input variables to be assessed. This is made possible through the application of a dynamic radius, which is altered depending on the number of input variables included in the model. Consequently, the distribution of the MHM statistic under independence is distributed the same as the HM statistic, allowing the use of the MHM to measure input importance in a stepwise fashion. A relative influence measure (RI) provides an indication of the importance of input variables, or combination of important variables.

These heuristic tools have been shown to be powerful, flexible, and effective. At a time when statisticians are questioning the use of models in statistical inference, while appealing for methods that recognise significant relationships, these tools are particularly relevant.

However, much work is required in order to provide a sound mathematical foundation for these methods. In addition much empirical work is needed in order to determine when they are unreliable. This paper has considered only polynomial functions and a few distributions. The impact of data with extreme kurtosis and skewness needs further exploration, although these effects can be reduced with transformation of the data prior to analyses. The effect of skewness and kurtosis in the MHM also requires attention. Additionally, the relationship between the one omitted MHM and the MHM needs to be further investigated to provide a versatile heuristic test for redundancy.

Finally, through the adoption of a algorithmic modelling approach, a powerful and versatile heuristic technique for exploratory data analysis has been created that has demonstrated its potential under numerous simulations.

Acknowledgement

This study is part of a larger body of work "Quantification of Individual Rugby Player Performance Through Multivariate Analysis and Data Mining", a PhD thesis funded by Eagle Sports, a division of the Eagle Technology Group, and a Graduate Research in Industry Fellowship from Technology New Zealand. This research was supervised at Massey University's Albany Campus by Associate Professor Denny Meyer with assistance from Dr Siva Ganesh. The numerous constructive comments for improving this paper provided by Professor Jeff Hunter are also appreciated. Additionally, Dr Paul Cowpertwait is thanked for his comments in preparing the final draft.

References

[1] Anton, H. (1995). *Calculus with Analytic Geometry.* (5th ed.). New York: John Wiley & Sons.

[2] Bracewell, P.J., Meyer, D.H., & Ganesh, S. (2001). Extracting Measures of Performance of Individual Rugby Players: Data Mining in Sport. *Proceedings of Artificial Neural Networks and Expert Systems Conference.* pp. 38-42. Otago University: Dunedin.

[3] Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science.* 16 (3) pp. 199-231.

[4] Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society, Series A.* 158 (3), pp. 419-466.

[5] Cheng, B. & Titterington, D.M. (1994). Neural networks: a review from a statistical perspective. *Statistical Science.* 9 (1), pp. 2-54.

[6] Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate Data Analysis with Readings* (4th ed.). New Jersey: Prentice Hall.

[7] Hand, D.J. (1999). Statistics and Data Mining: Intersecting Disciplines. *SIGKDD.* 1, pp. 16-19.

[8] Johnson, N.L. & Leone, F.C. (1964). *Statistics and Experimental Design: In Engineering and the Physical Sciences.* Volume 1. New York: Wiley.

[9] Krzanowski, W.J. (1988). *Principles of Multivariate Analysis, Part 1.* Oxford: Clarendon Press.

[10] Larson, H.J. (1982). *Introduction to Probability Theory and Statistical Inference.* (3rd ed.). Toronto: Wiley.

[11] Levine, D.M., Krehbiel, T.C., & Berenson, M.L. (2000). *Business Statistics.* (2nd ed.). New Jersey: Prentice-Hall.

[12] Smith, P.J. (1993). *Into Statistics.* Melbourne: Nelson.

[13] Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing.* London: Chapman & Hall.