A Non-linear Correlation Metric

Problem: ***Pearson correlation*** **is an effective and computationally simple method for identifying linear relationships between two variables. Through variable transformation it can sometimes be used to detect correlations where pre-determined functional relationships may exist. It is less useful in identifying the presence of unknown non-linear relationships. We propose an alternative technique that is both computationally simple and can detect the presence of non-linear relations where the functional forms are not known *a-priori.***

1. Correlation Dissected

Pearson correlation is often expressed in the following form:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Another way to think of it, however, is as a relative measure of the error of the data with respect to a 'powerful' estimator η (the linear fit) compared to the error of the data with respect to a less powerful 'weak' estimator, such as the mean.

$$r' = \sqrt{1 - \frac{\sum (y_i - \eta_i)^2}{\sum (y_i - \bar{y})^2}}$$

The concept of correlation can be extended to detecting non-linear relations when η is allowed to be a non-linear fit function – this is often called the *correlation ratio.*

2. Choice of Estimators

'Powerful' estimator (numerator)
- **Line Fit (Linear regression)**
- **Nearest neighbor (by 'dX')**
- **Interpolation between two neighbors**
- **K-nearest neighbors**

'Weak' estimators (denominator)
- **Mean**
- **Random Selection of 'Y' (Shuffle)**

Initial experiments suggest that using the nearest neighbor for the 'powerful' estimator and a random selection of Y for the 'weak' estimator yields a useful non-linear correlation metric that still provides similar results to Pearson on 'linear' data sets.

3. Algorithm

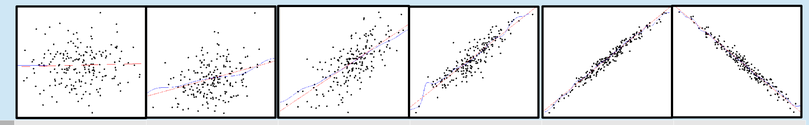
(Nearest Neighbor/Shuffle Estimators)

```
Function NLCorr( X(), Y() )

  N=Upperbound(X())
  X().Sortwith Y()
  Y()shuf = Y().Shuffle
  For I = 1 to N
    If I=1 then
      Yest = Y(2)
    Elseif I=N then
      Yest  = Y(N-1)
    Elseif X(I)-X(I-1) < X(I+1)-X(I) then
      Yest = Y(I-1)
    Else
      Yest = Y(I+1)
    Endif
    SSQNum += (Y(i) - Yest)^2
    SSQDen += (Y(i) - Yshuf(i))^2
  Next I
  Return Sqrt(1 - SSQNum/SSQDen)

EndFunction
```

4. Consistency with Pearson

A comparison of the new metric with Pearson for various cases of a line with superimposed Gaussian noise.

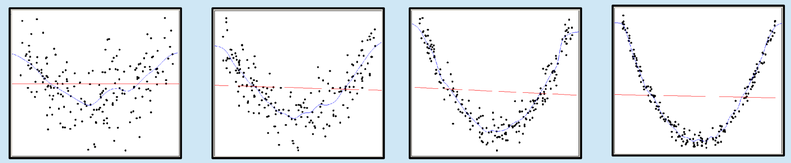
X = Gaussian(mean=0, sd=1)
Y = SNR * X + Gaussian(mean=0, sd=1)



Signal:Noise	0	0.5	1	2	5	-5
Pearson	0.034	0.324	0.697	0.896	0.982	-0.981
New Metric	*Undef.*	0.336	0.623	0.913	0.980	0.980

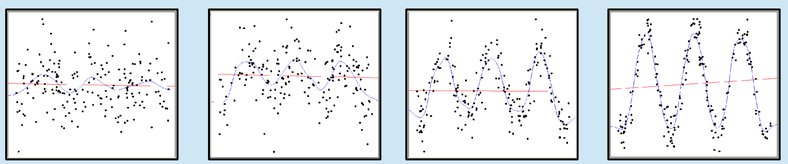
5. Nonlinear Relation Detection Performance

Quadratic Form



Pearson	-0.001	-0.048	-0.061	-0.022
New Metric	0.539	0.802	0.971	0.992

Periodic Form



Pearson	-0.029	-0.034	-0.007	0.068
New Metric	0.172	0.614	0.823	0.962

6. Next Steps?

Performance Envelope
- ***What functions can it detect?***
- ***# points needed?***
- ***What do values of metric mean?***

Explore Space of Estimators
- ***Which estimators work best?***
- ***Is there domain specificity?***

Develop Theoretical Basis ??

Presented at Poster Session of INFORMS 2012 General Meeting, Oct. 14-17, 2012, Phoenix AZ