Expanding the definition of multivariate correlation

W. Conley
Departments of Mathematics and Business Administration,
University of Wisconsin at Green Bay, USA

Abstract

The complexities of large scale data analysis, in our computer age, invite the development of new sophisticated statistics to help in this task. One entry into this arena is the CTSP multivariate correlation statistic. A five variable 49 line spreadsheet of data was analyzed using CTSP by Conley and the relationship was found to be linear. Presented here is a much larger example involving nine variables and 89 lines of data, where CTSP reveals a correlation that is nonlinear (exponential in this case). Specifically, nine columns (representing nine variables) of 89 lines of data are being analyzed to see if the variables they represent are correlated in some fashion (linear or nonlinear). Therefore, the data is read into the CTSP statistical analysis simulation program which is adjusted for nine variables and 89 lines of data. Then, using a generalization of the Pythagorean distance measure to nine dimensions, a shortest route connecting the 89 points in a closed loop tour is calculated. Then several random data sets of 9 x 89 size (in similar ranges to the original data) are generated and the shortest routes are calculated for them. In this case, the actual data had a much shorter shortest route than the random data's shortest routes. Therefore, statistically speaking it can be argued that the actual data is correlated in some fashion because it is following a pattern and hence the points are more compact (closer together in nine dimensional space) leading to a shorter shortest route. The relationship is exponential in this case. This expanded view of correlation (linear or nonlinear) can complement the standard linear analysis Anderson currently uses. Additionally, a second example involving eight variables is presented for comparison purposes.

Keywords: multivariate correlation, CTSP statistic, shortest route statistical test, linear and nonlinear analysis.

WIT Transactions on Modelling and Simulation, Vol 46, © 2007 WIT Press
www.witpress.com, ISSN 1743-355X (on-line)
doi:10.2495/CMEM070191

1 Introduction

The new CTSP correlation statistic was developed in 2002 to help discover multivariate relationships whether they be linear or nonlinear. The CTSP is short for (an acronym) correlation using the travelling salesman problem. The idea in two dimensions can be illustrated by thinking of *X, Y* pairs of points that are following the shape of a parabola when graphed. These points will have a shorter shortest route connecting them, than the shortest route connecting the same number of random points (in the same range as the parabola points).

A three dimensional example would be *X, Y, Z* triples on a flat plane going through the points (0, 0, 0), (100, 100, 100), (100, 0, 100), and (0, 100, 0) in the region 0 to 100 for all three variables. The shortest route connecting these *X*, *Y*, *Z* triples (say 75 of them, for example) would be much shorter than the shortest route connecting 75 random triples (in the same range) that are not following a pattern.

This can be exploited statistically to show a correlation (or lack thereof) when analyzing multivariate data. The name CTSP statistic comes from the mathematically famous TSP problems (so called travelling salesman problem) of finding a shortest route to connect *n* points in a closed loop tour.

This analysis presented here uses the multi stage Monte Carlo optimization MSMCO TSP algorithm for the statistical analysis. Let us look at an example.

2 Numeric example one

Researches believe that the nine variables represented by the nine columns of data in Table 1 are correlated in a linear or nonlinear fashion.

They think that the first eight variables are perhaps driving the ninth variable (X_9 whose data is in Column 9). Therefore, to test the null hypothesis of no correlation between the nine variables (X_1, $X_2 \ldots X_9$) versus the alternative hypothesis of correlation, the 89 lines of 9 columns of data are read into the MSMSCO TSP algorithm program adjusted for 89 lines of data and nine variables. The Pythagorean theorem distance measure is used after expanding it to nine dimensional distance calculations.

A few seconds to a minute of computer run time (on a desktop PC) yielded a shortest route of total distance 5,776.308 (see Table 2 for the route read left to right) connecting the 89 nine dimensional points in a closed loop tour.

Then four sets of 89 lines of nine columns of random data (in similar 0-100 ranges as the actual data) were read into the MSMCO TSP algorithm. Their shortest routes were calculated to be 6786.314, 6518.579, 6600.060 and 6410.455. Therefore,

$$CTSP = A / B = 5776.308 / (6518.579 + 6600.060) / 2 = .88 \tag{1}$$

where A is the shortest route distance for the actual data and B is the median of the four random data shortest routes. Now taking the 3 x 4 = 12 A_i/B_i quotients using all combinations of the four random shortest routes, we see that their range

Table 1: The data.

Point	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	73	12	46	59	41	24	95	96	3.6
2	66	5	69	49	94	18	100	21	4.9
3	9	17	18	58	85	85	62	29	3
4	40	41	95	91	8	68	18	23	3.6
5	40	70	58	36	8	35	30	72	2.7
6	5	51	49	44	37	33	72	23	2
7	0	50	89	77	83	13	64	52	5.1
8	63	99	29	13	27	46	30	82	3.7
9	61	56	100	14	91	1	25	31	3.4
10	28	70	57	30	86	31	47	100	6
11	31	100	11	99	73	76	5	82	7.5
12	50	60	16	79	11	86	100	43	5.8
13	65	23	77	39	65	36	51	87	5.7
14	32	13	100	90	18	33	38	29	2.8
15	100	10	95	60	24	24	33	88	5.3
16	40	94	94	30	70	46	65	82	10.7
17	52	20	67	14	3	99	41	94	3.8
18	68	89	29	2	62	38	92	100	7.7
19	17	90	72	40	16	49	40	4	2.3
20	71	48	66	71	36	1	99	53	5.8
21	78	56	57	44	62	82	39	19	5.5
22	70	47	75	25	100	70	20	20	5.1
23	46	0	37	70	88	100	34	20	5.4
24	58	39	22	97	68	57	42	52	5.4
25	65	61	93	5	31	33	29	81	4
26	88	90	62	90	86	17	69	42	12.9
27	51	28	20	77	75	64	29	13	2.9
28	57	74	36	71	100	74	57	91	14.6
29	12	77	41	72	63	48	36	92	5.7
30	33	3	32	63	88	57	81	73	5.2
31	100	97	98	32	45	85	0	70	11.2
32	4	19	69	33	100	9	63	88	3.6
33	13	23	25	42	27	96	21	92	2.5
34	38	26	21	100	1	26	40	5	1.3
35	77	57	72	19	97	91	91	44	13.3
36	52	63	10	18	22	68	100	17	2.7
37	55	19	35	41	32	37	35	82	2.4
38	11	41	42	89	46	5	35	57	2.3
39	36	20	10	87	4	86	58	95	3.9
40	65	66	15	23	0	3	16	67	1.3
41	50	58	3	23	100	97	73	92	8.8
42	14	79	29	88	56	72	34	75	5.9
43	84	44	35	21	24	54	2	57	2.2

Table 1: Continued.

Point	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
44	28	32	44	80	88	13	21	10	2.1
45	51	40	100	22	79	61	74	99	11.2
46	51	94	2	4	6	100	27	88	3.3
47	56	2	31	60	74	12	37	1	1.5
48	35	88	5	79	1	36	98	48	3.8
49	83	95	27	46	32	90	40	59	7.2
50	40	4	85	50	56	99	85	42	6.6
51	90	43	73	52	76	92	74	36	12.1
52	33	23	53	2	61	34	74	30	2
53	55	100	9	79	70	35	0	3	2.7
54	45	51	3	83	48	63	49	37	3.4
55	43	100	31	87	31	34	25	55	4.3
56	96	7	13	16	81	27	66	44	2.7
57	13	43	37	41	19	59	77	35	2.2
58	60	58	14	90	93	65	75	74	11.4
59	18	28	6	97	1	76	22	81	2.3
60	29	29	15	32	60	65	64	6	1.8
61	57	63	75	91	81	83	86	80	22.9
62	43	57	60	53	90	62	90	86	12.6
63	57	95	50	82	35	100	90	33	12.7
64	29	100	45	66	92	63	78	100	16.3
65	81	100	1	18	99	6	90	62	6.4
66	41	31	33	16	13	35	90	8	1.4
67	93	47	79	48	100	76	60	91	19.2
68	45	27	70	65	79	92	56	26	6.6
69	7	95	77	30	12	10	56	36	2.2
70	31	72	51	73	88	75	82	25	8.8
71	11	20	19	43	82	100	95	58	5.1
72	95	7	31	15	26	12	51	89	2.3
73	62	23	3	17	71	60	42	20	1.8
74	66	74	60	10	41	52	46	17	3.1
75	88	53	46	9	65	94	82	68	9.4
76	8	10	92	39	37	83	55	3	2.3
77	66	43	81	93	39	27	21	95	6.8
78	31	60	79	37	33	75	46	25	3.6
79	62	40	100	48	91	6	55	66	7
80	29	83	24	28	56	80	77	63	5.6
81	35	36	9	39	49	43	28	96	2.4
82	17	40	81	23	36	11	55	14	1.5
83	44	56	46	84	91	100	19	10	6.1
84	1	20	100	20	34	16	2	74	1.4
85	2	56	0	20	15	12	35	100	1.1
86	98	77	100	19	48	8	35	35	4.8

Table 1: Continued.

Point	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
87	72	56	4	39	97	15	60	8	2.8
88	39	4	94	98	60	2	100	74	7.2
89	97	16	59	100	81	95	52	89	18.5

Table 2: The route.

43	17	33	39	59	54	24	58	28	64
29	42	11	55	53	27	83	23	68	50
76	78	19	69	82	6	57	66	36	48
12	63	70	26	79	9	86	74	52	2
1	20	88	7	32	10	16	18	65	87
47	44	38	34	4	14	77	15	13	45
67	89	61	62	30	71	3	60	73	56
72	37	81	85	40	5	8	46	49	80
41	75	35	51	21	22	31	25	85	43

is 6410.455/6786.314 = .945 to 6786.314/6410.455 = 1.059. This range contains the vast majority of the sampling distribution of CTSP under the null hypothesis of no correlation. However, from eqn (1) our CTSP = .88, which is less than this range. Therefore, the null hypothesis of no correlation can be confidently rejected. The nine variables are correlated and the equation

$$X_9 = .1666\exp((X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8)/125) \tag{2}$$

fits the data fairly well.

It should be noted that the range of the actual X_9 variable values is somewhat smaller (0-25) rather than 0-100 for the other eight variables. However, if one evaluates eqn (2) with 100 for X_1 through X_8, X_9 is 100. Therefore, the restricted range of X_9 was a function of the correlation, so random sampling all nine ranges in the area 0-100 seems appropriate for the randomness comparison.

However, if the researcher or engineer believes that there is scientific justification for having different ranges for the variables, they can be used for the random data sets generation and subsequent testing. The engineer or scientist will know what ranges are believable and appropriate for the application at hand.

The central point is that a shorter shortest route (than randomly generated data's shortest routes) will indicate that the data is following a pattern and is hence indicating that the variables are correlated.

3 Example two

Researchers studying the eight columns of $n = 49$ lines of data in Table 3 want to test the hypothesis of no correlation between the eight variables represented by the eight columns of data.

Therefore, the 8 x 49 array is read into the MSMCO TSP algorithm. A less than one minute computer run produces a shortest route of total distance A = 3571.170 (see Table 4) for analysis.

Table 3: Example two data.

Point	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	55	10	89	17	33	50	100	45
2	27	38	86	67	50	12	43	47
3	82	98	61	13	51	81	27	61
4	72	74	4	93	100	83	12	56
5	37	30	57	70	15	90	86	17
6	4	100	50	6	3	66	47	4
7	94	50	82	39	19	79	84	3
8	92	99	83	23	21	27	77	4
9	46	100	82	1	4	12	38	10
10	0	18	7	85	24	12	64	62
11	85	61	21	100	100	17	23	89
12	74	70	70	100	100	94	57	18
13	38	84	59	17	98	9	11	87
14	57	22	36	19	2	47	49	82
15	54	26	92	3	26	25	19	99
16	40	52	60	97	38	96	20	93
17	71	56	10	82	79	52	22	87
18	88	57	75	42	2	47	56	15
19	100	37	45	42	55	52	26	3
20	47	8	41	67	36	29	59	87
21	16	45	20	25	71	99	53	62
22	36	63	46	28	67	40	61	3
23	21	100	13	94	22	20	21	52
24	29	51	41	100	39	36	75	21
25	59	85	83	41	27	1	79	40
26	54	72	49	62	32	58	55	15
27	65	40	71	65	69	34	3	27
28	10	45	73	30	49	21	98	84
29	23	65	76	80	5	44	6	57
30	48	15	80	28	80	11	84	89
31	55	42	70	40	28	17	56	76
32	66	7	82	100	98	73	80	33
33	20	8	71	37	7	20	55	39
34	74	63	26	9	33	4	44	14
35	80	64	81	15	99	38	72	86
36	26	58	5	35	56	17	21	75
37	24	40	26	2	90	100	3	53
38	9	54	31	18	33	41	49	26
39	64	88	54	56	61	68	69	8

Table 3: Continued.

Point	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
40	78	21	56	37	35	86	92	92
41	32	53	34	39	28	29	25	99
42	61	88	82	10	61	61	23	7
43	89	44	92	48	91	6	58	83
44	61	48	54	100	8	88	34	44
45	98	100	55	29	97	1	83	62
46	32	96	70	87	95	0	89	8
47	100	93	47	73	17	75	81	17
48	85	47	39	25	5	14	33	71
49	13	88	96	98	39	88	68	37

Table 4: Example two route.

45	35	43	30	28	31	20	10	24	49
5	44	16	29	23	11	17	4	12	32
46	25	8	18	7	47	26	39	22	38
6	9	42	3	37	21	40	1	33	2
27	19	34	48	14	15	13	41	36	

Then four sets of random data of 8 x 49 are read into the MSMCO TSP algorithm. Their shortest routes turn out to be 3346.265, 3473.206, 3564.802 and 3363.000. Note that the real data shortest route distance (3571.170) is comparable to these and will not produce a relatively smaller CTSP. Therefore, the null hypothesis of no correlation between the variables cannot be rejected.

4 Expanded definition of correlation

We are now prepared with the CTSP multivariate statistic (for the computer age) to expand our definition of multivariate correlation to the discovery of any type of pattern, whether it be linear, nonlinear and even non-functional (like a mathematical relation). Additionally, if one wishes to assume and/or test for various underlying distributions, CTSP can help in that accommodation also.

5 Conclusion

Much of the standard multivariate analysis used today assumes sampling from multivariate normal distributions and looking for linear relationships. However, with CTSP we can go beyond this and look for any type of relationship that is revealed by a relatively shorter shortest route through the *k* dimensional space (for *k* variables) when compared with similar random data.

The various TSP algorithms can help trucking companies and transportation entities in their deliveries to their valued customers. However, the TSP analysis when combined with the CTSP multivariate correlation statistic, can do much more than deliver products and services to customers. It can also deliver sophisticated multivariate statistical correlation analysis on spreadsheets of data that researcher and engineers encounter on an almost daily basis.

References

[1] Conley, W.C., Multi stage Monte Carlo optimization applied to systems of integral equations. Proc. of the 15th Int. Conf. on Boundary Element Technology, ed. C.A. Brebbia-WIT Press, Southampton, pp 75-86, 2003.

[2] Anderson, T.W. *Multivariate Statistical Analysis*, 3rd edition, John Wiley Inc, New York, 2003.