Improvements on the Development of Correlated Input Variables for Monte Carlo Simulation

Douglas Henke

Abstract: In this paper, the development of model inputs with specified correlation is explored. Input variable correlation is an influential driver of final cost risk distributions; especially so for highly positively or negatively correlated inputs. If not captured appropriately between inputs, significant errors in resultant cost risk distribution will occur. In general, the further away from a 50% confidence level cost value, the greater the error will be when input data does not reflect accurate correlation. The widely adopted Iman-Conover (IC) Method for inducing desired rank correlation on a multivariate input for modeling by Monte Carlo simulation is reviewed. The IC method culminates in the re-ordering of the values of each input vector such that the resultant correlation of the vectors is close to the desired correlation. This paper provides insights into how the IC Method, devised as a method for inducing a desired rank correlation, can be equally if not more powerful for inducing desired Pearson product-moment (linear) correlation on inputs. Spearman's rank correlation and linear correlation values that result from the IC Method are compared to the desired correlation values ranging from -1 to 1. Insights into the mechanics of the algorithm are presented in order to provide a richer understanding of the process and to inform aspects of work when the algorithm is employed. Extending this IC method to an iterative process described in this paper shows that the resulting set of variates would more accurately reflect the desired correlation in all cases for the calculated linear correlation. Conversely, for highly skewed distributions, the iteration process resulted in increasing the error of the calculated Spearman's rank correlation. The iteration process is explained with examples to illustrate improved linear correlation accuracy for both symmetric and highly skewed distributions.

Most cost risk and statistical software platforms on the market today permit the user to define the distribution type of each input variable as well as the correlation between these variables. Common forms of quantifying correlation include Pearson product-moment (linear) correlation and Spearman's rank correlation. Using linear correlation with input variables that have outlier or clustered data; or have unusual forms of distribution may not appropriately quantify the relationship between variables (Tamhane, 2000). For those input relationships that are not monotonic or which depart significantly from a linear relationship, Spearman's rank correlation metric may be the more appropriate measure. Accurately quantifying the correlation of input variables of the phenomenon, stochastic process or estimate being modeled is a well-recognized and necessary element of increasing the accuracy and realism of resultant risk distributions and cost estimating results. Inaccurate correlation values exacerbate cost risk distribution errors that are further away from 50% confidence level. In the simplest terms, the final cost risk variance is inflated or contracted with inaccurate input correlation. It is important to note that the IC method culminates in the reordering of the original set of input variates based on the rank order of a resultant linear transformation they undergo. This linear transformation is based on the desired correlation matrix of the inputs; which may be defined as either linear or rank correlation. In either case, the efficacy of the IC method is considered in this paper by how close the resultant linear and rank correlation is to the specified desired correlation. The ability to

perform this work independent of embedded software features can be useful; as well as beneficial to understanding any limitations of the process that apply in practice.

Richard L. Iman and W. J. Conover published a paper in 1982 entitled *A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables*. Consider, for example, M input vectors to represent M variables in a cost model. Each of the M vectors contain N values for a model run to be performed with N iterations. Assume the correlation between all M input vectors is known. Iman and Conover devised a powerful and effective methodology that culminates in the reordering of values of all but the first input vectors such that the resulting correlation of these reordered vectors is very close to the desired correlation. Once reordered, these vectors are then used as appropriate inputs.

The following summarizes the steps of the IC Method. Of note, Stephan J. Mildenhall provides an expanded treatment of the IC Method in his paper, *Correlation and Aggregate Loss Distributions With An Emphasis On The Iman-Conover Method* presented in 2005.

Assume a model requires M input variables; each with N values for a model run of N iterations. Let [A] represent the set (or matrix) of input variables of size N (rows) by M (columns). For illustrating the IC method, the distribution type of each input is not significant. In this paper, the inverse of the cumulative distribution function of each input variable is known. This is used to determine the elements of each input vector as van der Waerden scores where the i^{th} element of the j^{th} variable in [A] is initially determined as follows.

$$A_{i,j}=\Phi_j^{-1}\left(\frac{i}{N+1}\right)$$

The desired correlation matrix must be positive semi-definite to enable a tractable decomposition. Define the desired input vector correlation matrix as [S]. It is of dimension M by M. Zero-mean and scale each of the input vectors such that the variance is 1. This transforms [A] to what is now defined as [X]. At this point, each vector is ordered from lowest to highest rank by virtue of their derivation as van der Waerden scores. The IC method requires linear independence of the input vectors. To invoke this linear independence, randomly permute each of the input vectors. This is not necessary for the first vector as it is unaffected by the IC method. Now define a new matrix [E] as the covariance of [X].

$$[E]=\frac{1}{N}\cdot[X]^T[X]$$

Since the vectors of [X] have zero mean and variance of 1, [E] represents the linear correlation matrix of [X] and should have low absolute values off-diagonal due to the prior random permutation of the vectors that induced linear independence. The next step is to apply a Cholesky Decomposition to [E].

$$[E]=[F]^T[F]$$

[F] is an upper triangular matrix and represents the square root of [E]. As a square symmetric matrix, [E] can be transformed to $[L][D][L]^T$ where $[L][D^{(.5)}]$ is the lower triangular matrix of the decomposition and equals $[F]^T$. Cholesky decomposition would fail with any of the diagonals of [D] less than zero (i.e. Not positive semi-definite). A significant facet of this decomposition can be shown as follows:

$$[F]^{-T}[F]^T[F][F]^{-1}=[I]$$

$$\frac{1}{N}\cdot[F]^{-T}[X]^T[X][F]^{-1}=[I]$$

This shows that pre-multiplying the covariance matrix of [X] by the inverse of the $[F]^T$ and post-multiplying by the inverse of [F] give the identity matrix.

To incorporate the desired correlation, a similar Cholesky decomposition is performed on the desired correlation matrix [S] where

$$[S] = [C]^T[C]$$

Now consider inserting the identity matrix between the two [C] matrices without effect.

$$[S] = [C]^T[I][C]$$

It was shown above that pre and post multiplying $[X]^T[X]/N$ by the inverses of the Cholesky decomposition of [X] results in the identity matrix, [I]. Replacing [I] in the equation above with that identity gives the following.

$$[S] = \frac{1}{N} \cdot [C]^T[F]^{-T}[X]^T[X][F]^{-1}[C]$$

This shows us that multiplying the randomly permuted starting vectors of [X] by $[F]^{-1}[C]$ results in a set of vectors with a calculated linear correlation that exactly matches the desired correlation, [S].

For convenience, the new transform matrix [T] is defined as follows.

$$[T] = [F]^{-1}[C] \text{ and define } [X'] = [X][T], \text{ then}$$

$$\frac{1}{N} \cdot [T]^T[X]^T[X][T] = \frac{1}{N} \cdot [X']^T[X'] = [S]$$

However, since each vector (column) of [X'] is a linear combination of the vectors of [X], the original distributions are no longer preserved. That is, the desired correlation is achieved, but no longer with the original values that comprised the vectors of [X]. It is noteworthy that the variance of each vector of [X'] is 1 since the covariance of the [X'] matrix is [S].

At this point, Iman and Conover capitalize on the connected relationship between Spearman's Rank correlation and linear correlation where the rank correlation is the linear correlation of the ranks of the values that comprise the vectors. As the next step in the IC method, the elements of each of the variates of [X] are reordered to have the same rank ordering of the corresponding vectors of [X']. This generally results in a correlation of the re-ordered vectors that is close to the desired correlation prescribed in [S]. Since these data were zero-meaned and rescaled to have a variance of 1 for the IC method, one needs to simply reverse the process to regain the original distribution that is now re-ordered for the desired correlation.

By way of example, consider a matrix [A] that contains three input vectors of 30 elements each (not shown). Each are sampled as van der Waerden scores from a normal distribution with zero mean and are normalized to have a variance of 1. With a zero mean and variance of 1, [A] becomes [X], and is of dimension 30 X 3. In this case, each input vector of [X] is identically distributed.

The second and third vector are randomly permuted so that columns of [X] are linearly independent. Once accomplished, determine the linear correlation matrix of [X], defined as [E].

Permuted $[X]^T[X]/N$ = Correlation = [E]		
1	0.014629	-0.13703
0.014629	1	0.144805
-0.13703	0.144805	1

The upper Cholesky decomposition of this correlation matrix, [E], and its inverse are as follows.

[F]		
1	0.014629	-0.13703
0	0.999893	0.146825
0	0	0.979624

$[F]^{-1}$		
1	-0.01463	0.1420773
0	1.000107	-0.149895
0	0	1.0207994

A positive-definite desired correlation matrix, [S], is developed for this example and its upper Cholesky decomposition are shown as follows.

[S]		
1	-0.7	0.5
-0.7	1	-0.2
0.5	-0.2	1

[C]		
1	-0.7	0.5
0	0.7141428	0.210042
0	0	0.8401681

We now have the information to produce the transform matrix [T] where $[T] = [F]^{-1}[C]$.

[T]		
1	-0.71045	0.616296
0	0.714219	0.084128
0	0	0.857643

The next step is to multiply [X] by [T]; resulting in [X']. The resulting covariance matrix of [X'] ($1/N * [X']^{T}[X']$) now equates to [S] precisely.

$[X']^{T}[X'] /N = [S]$		
1	-0.7	0.5
-0.7	1	-0.2
0.5	-0.2	1

As discussed before, [X'] is a linear combination of the vectors of [X] and all but the first vectors of [X'] now have a different distribution than that of the [X] vectors. Per the IC method, each vector of [X] is reordered to have the same rank order as the corresponding vector of [X'].

The following shows the Resultant linear and Spearman's rank correlation matrices based on the reordered vectors of [X]:

Spearman Rank Correlation		
1	-0.74	0.44
-0.74	1	-0.25
0.44	-0.25	1

Pearson Linear Correlation		
1	-0.72	0.53
-0.72	1	-0.23
0.53	-0.23	1

The resultant rank correlation of [X'], and by process, the rank correlation of the reordered vectors of [X] are close to, but do not match the desired correlation, [S], precisely. Similarly, the linear correlation of the reordered elements of [X] is close to, but does not match [S] precisely. As will be seen, larger values of N, which are typically employed in practice, result in much closer alignment with the desired correlation matrix.

Exploring the IC Method

In the following, the IC method is applied to two-vector input matrices (i.e. two input variables) of various sizes (N). This facilitates a more tractable analysis and the concepts apply to M dimensional input matrices. The difference between the desired correlation as prescribed in [S] and resultant correlation (both linear and rank correlation) from the reordered elements of [X] will be considered; as well as means to reduce this difference.

A 4 X 2 input matrix is considered first. From a Euclidean space construct, [X] spans 4 dimensions (N) and is of rank 2 (M=2); provided the vectors are linearly independent. There are 4 factorial (4! or 24) different permutations of the second vector that is reordered in the process. When the values of each input vector are viewed as coordinates, each of the permutations of the second vector occupies a discrete point in N-dimensional space and each also has its own

correlation value with the first vector. In this case of input vectors, the first is derived from a normal distribution and second is a skewed distribution. There are 24 unique linear correlation values and 11 unique rank correlation values with the first vector from all possible permutations of the second vector. These permutations represent a constellation, as it were, of 24 discrete points at a Euclidean distance of the square root of N from the origin due to a variance of 1.

Consider a randomly permuted second vector of [X] and the notion that the IC method can be applied for any desired correlation value between the vectors of [X]. Recall that [X'] = [X][T] gives the exact desired correlation, but no longer possesses the original distribution of values. There is a continuum of the second vectors of [X'] associated with each value within the range of desired correlation values from -1 to 1 since. The second vector of [X'] is a linear combination of the two vectors of [X] as determined by [T]. This continuum of [X'] spans a plane (or arc) of the continuous N dimensional space subtended by the vectors of [X]. This subspace is referred to herein as the **[X] subspace**. In the last step of the IC method, the re-ordered second vector of [X] becomes one of the 24 possible permutations (and one of the 24 possible linear correlations) that is near the [X] subspace because the re-ordering is based on the rank of the second vector of [X']. This does not allow for a permutation that may result in a correlation value closer to the desired correlation but is in a region away from the arc of the [X] subspace. Accordingly, the initial random permutation of the vectors of [X] pre-determine the [X] subspace and possible resultant correlation values.

Similar to the vectors of [X] and [X'], consider now all other possible points in the continuum of N dimensional space that are a distance from the origin of square root of N and whose mean is zero. This continuous collection of points is comprised of bounded and connected regions of all possible coordinate rank orders. Since there are N factorial possible permutations of a vector (or set of coordinates) of dimension N, there are N factorial regions whose coordinates have the same rank order. Each of these regions is referred to as a **rank order region**. Each rank order region is defined by the rank order of the values in the vector that map to that region. For example, where N = 4, the vector comprised of the following values in the order shown is 1423, {-0.780, 1.724, -0.575, -0.439}. Provided that all values are unique in a vector of [X], each permutation resides as a discrete point within each of these regions. The IC method determines in which regions the re-ordered vectors of [X] reside by virtue of the rank order of vectors in [X'].

To illustrate these concepts, principal component analysis was employed for the dimensional reduction of all possible permutations of the input vector that spans 4 (N) dimensional space. The points were mapped to 3 dimensions and are shown below from a perspective angle with the background walls and floors aligning with the Cartesian coordinates. There was no loss of information in the dimensional reduction because each vector has a mean of zero and variance of 1. Hence, each vector represented as point in 3-D space is equidistant from the origin. Further, all possible vectors of N equals 4, of zero mean and variance of 1 occupy the surface of a sphere of radius 2 in this 3-D space. In general, all possible real numbered vectors of dimension N with zero mean and constant variance will occupy an N-2 dimensional subspace.

Figure 1 below is a graphic that shows the boundaries of all 24 rank order regions. Figure 2 shows only the front facing rank order regions for illustrative purposes. The points shown represent permutations of a single vector. In this example, the vector is {-0.710, -0.575, -0.439, 1.724}. Of note, these permutations derived from a skewed distribution are located near a corner of each region. Each of the 24 permutations occupies one point in each of the rank order regions.

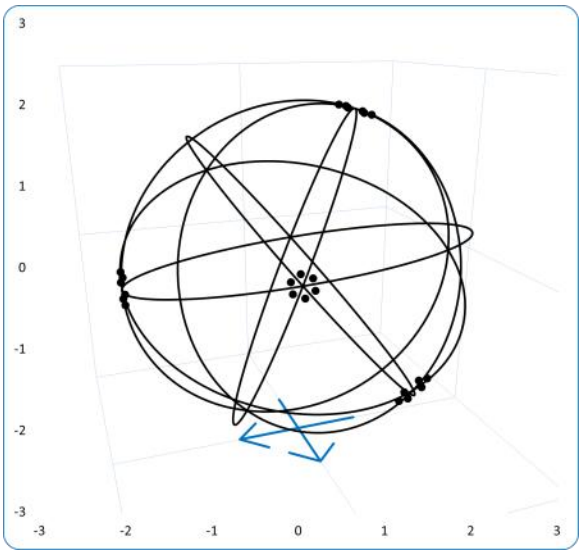

Figure 1

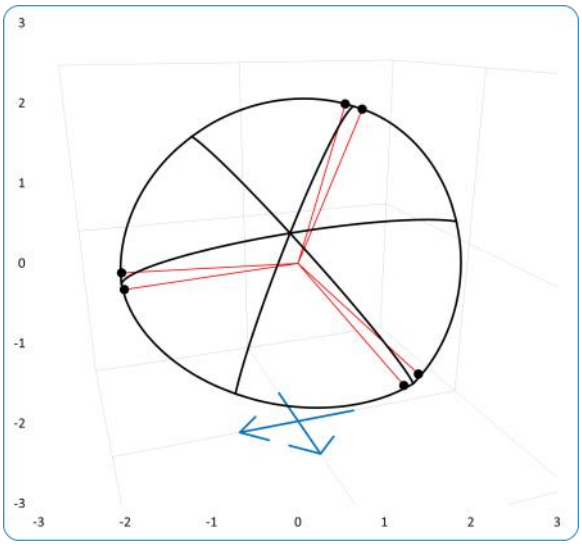

Figure 2

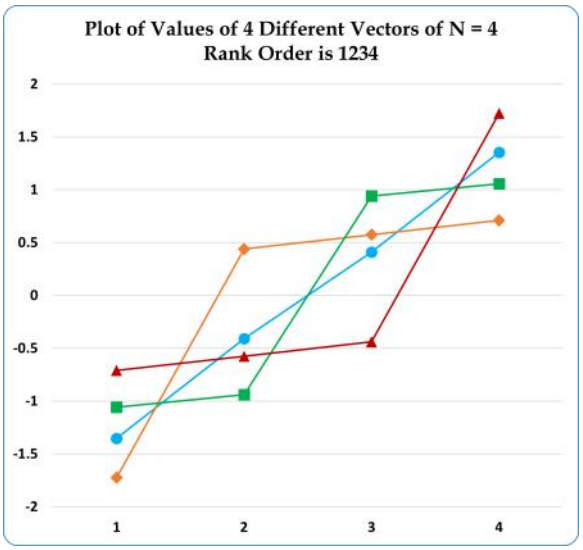

Figure 3

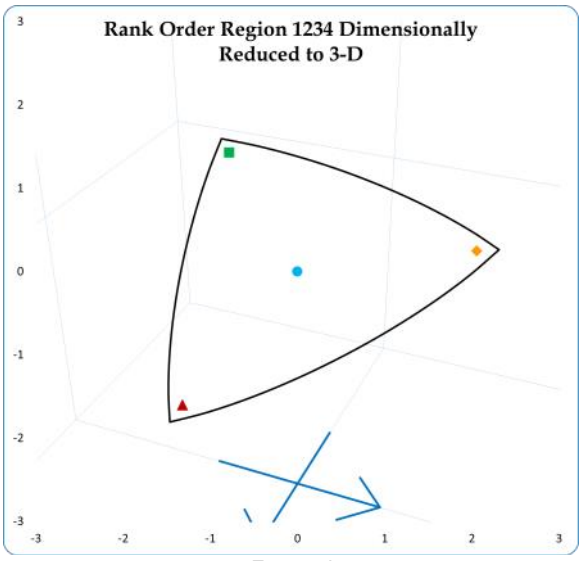

Figure 4

Figure 3 is a plot of four different vectors of N = 4 with rank order 1234. Figure 4 shows where each of these map to a point in the rank order region 1234. It is then shown where each of these map to a point in the rank order region 1234. Of note: the point in the center is the geometric center of the corners of the rank order region. It was determined that the four values of the vector associated with this point are very close to those values derived as van der Waerden scores from a normal distribution.

Using the construct developed above, the following is a graphical illustration of the IC method starting with [X] comprised of 2 vectors of N = 4. The first vector is derived from van der Waerden scores from a normal distribution and the second vector, as shown above, is comprised of {-0.439, 1.724, -0.575, -0.780} with starting rank order 3421. The desired correlation is -0.2225. The linear correlation of the starting vectors of [X] is -0.326; reflecting an absolute error from desired correlation of .103.

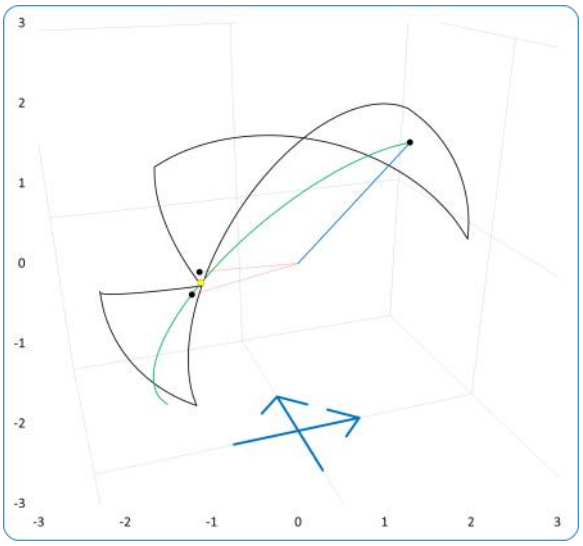

Figure 5

- The blue vector emanating from the center represents the first vector in [X]. As a vector derived from van der Waerden scores from a normal distribution, this point is in the center of the region with rank order 1234.
- The second vector of [X] is represented as the lower of the two red vectors. It occupies the lower left rank order region shown of rank order 3421.
- The green curve is the locus of all points mapped from the second vector of [X'] through correlation values ranging from -1 to 1. Recall that [X'] = [X][T] in the IC method where the second vector of [X'] is a linear combination of the vectors of [X]. This is the [X] subspace.
- The yellow marker is the point on the locus of points that satisfies the desired correlation value of -0.2225 exactly. This solution of the second vector of [X'] is {-0.607, 1.732, -0.546, -0.578} and is of rank order 1432. This point lies in a different rank order region than the second vector of [X].
- In the IC method, the second vector of [X] is then reordered to have that same rank order as the second vector [X']. This is depicted as the upper of the two red vectors and is merely the permutation of the second vector of [X] with the new desired rank order.
- The resultant correlation of the reordered second vector of [X] with the first vector is -0.175; an absolute error from desired correlation of 0.048.
- The following shows a clearer view of the region of interest.

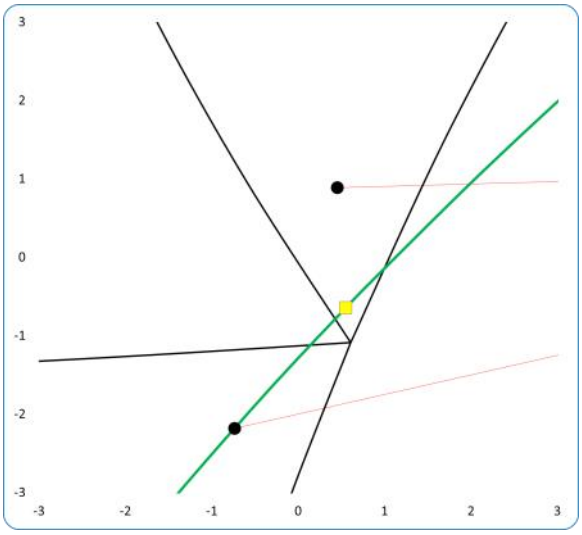

Figure 6

[X] Subspace

To demonstrate the constraint of the reordered second vector or [X] to points near the [X] subspace, the IC method was performed on the 4x2 input matrix over a range of desired correlation values from -1 to 1 in increments of .00125 (1/800). The first and second vectors of [X] applied here are the same as those used in the above illustration of the IC method. Now, define the **resultant correlation error** as the difference between the resultant correlation value and the desired correlation value (input to the [S] correlation matrix). The following graphs show both the resultant correlation and the resultant correlation error for both linear and Spearman's rank correlation measures.

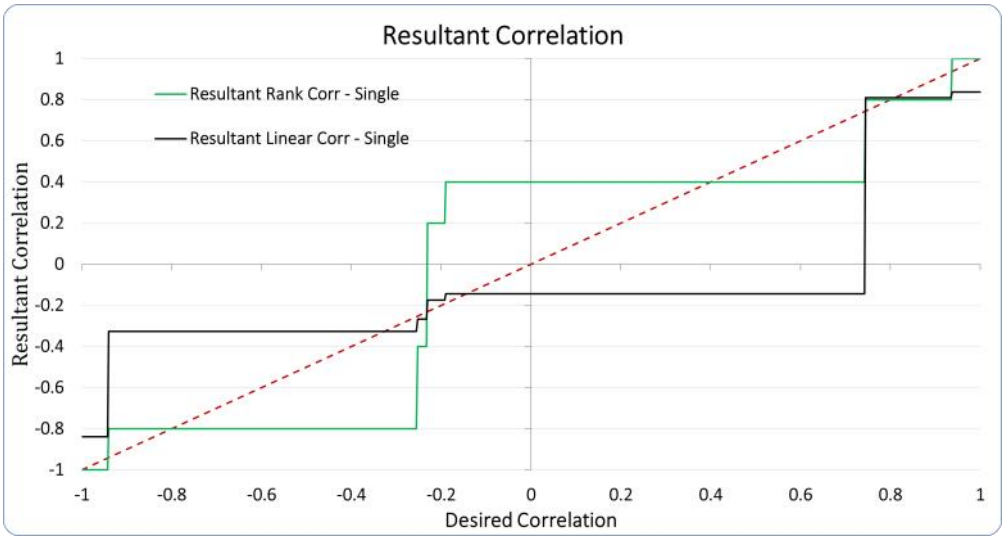

Figure 7

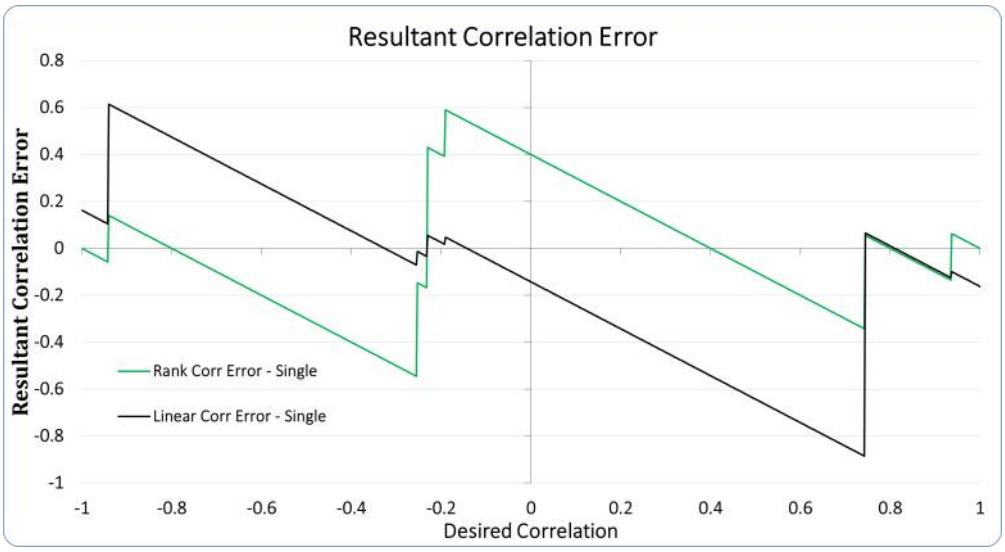

Figure 8

With only 4 values in each vector, a large error across much of the correlation range from -1 to 1 would not be unexpected. However, more noteworthy as it relates to the [X] subspace constraint is that of the possible 24 linear and 11 rank correlation values associated with all possible permutations of the second vector, there were only 7 distinct correlation values that resulted from the desired correlation values evaluated between -1 and 1.

Correlation with N = 10, 30, 100 and 1000

Resultant linear and rank correlation were evaluated in four other cases where M=2 and the number of values in each vector was 10, 30, 100 and 1000. As before, the desired correlation ranged from -1 to 1 by .00125. These starting vectors were also derived as van der Waerden scores from a normal distribution. The starting vectors of [X] remained the same through the range of desired correlation values. The following shows the resultant correlation and resultant correlation error for N = 100. The results from

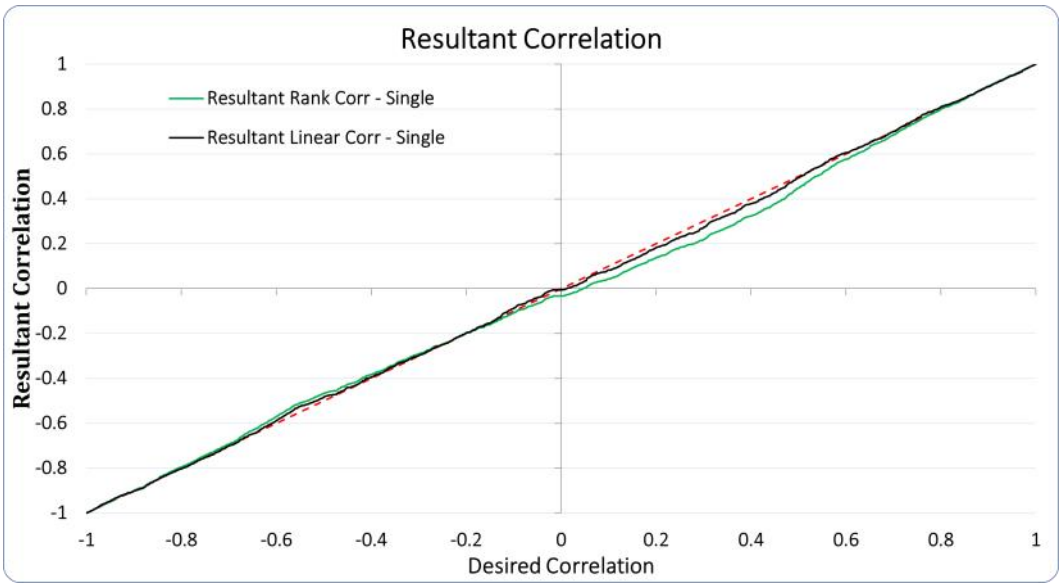

Figure 9

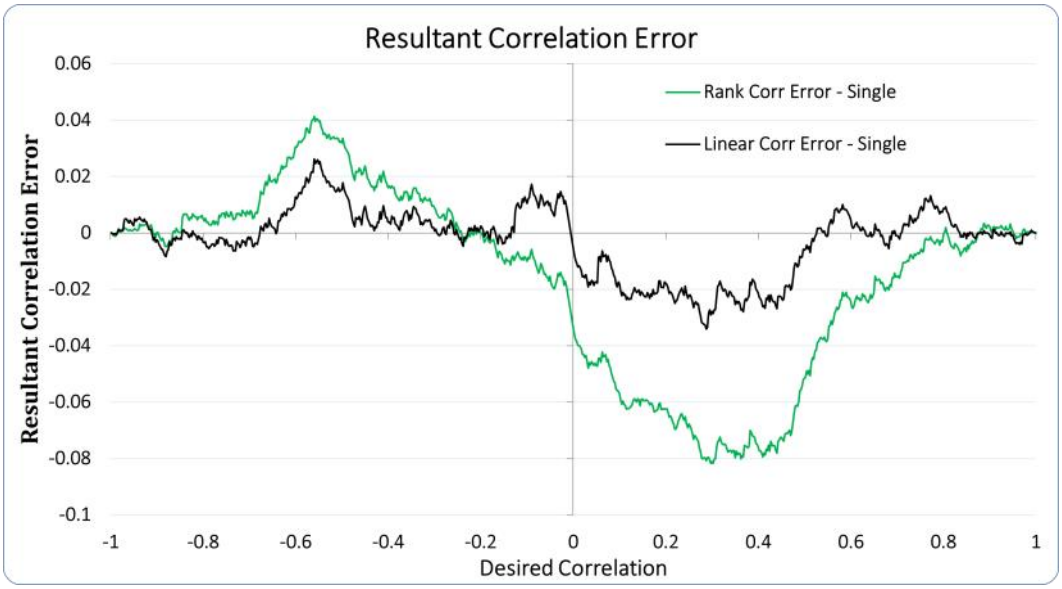

Figure 10

the other cases are available as supplemental material for this paper.

Observations:

- In the case of N = 10, there are 45 distinct resultant correlation values that are derived by the IC method. This is based on the same second vector in [X] for all desired correlations between -1 and 1 (evaluated at increments of 1/800). There are 10 factorial (3.63 million) possible permutations of the second vector of [X]. Each has a linear correlation value with the first vector. In this case, there are hundreds of thousands of unique linear correlation values associated with the 3.63 million different permutations of the second vector; yet only 45 resultant correlation values are revealed across the range of correlations evaluated with the IC method in this example; emphasizing the resultant correlation's constraint to the [X] subspace.

- A **global error** is defined as the root mean square of each of the 801 error values for the linear and rank resultant correlation. The following table shows those results. The linear correlation global error is less than rank correlation global error for each case of N. Heuristically, the notion that there are more possible linear correlation values associated with all possible permutations than there are rank correlation values would support a lower global error with linear correlation. Notably the ratio of global errors (i.e. Rank divided by linear global error) increases significantly with N.

	Global Error	
N	Linear Correlation	Rank Correlation
5	0.13913	0.14355
10	0.06457	0.07311
30	0.03749	0.05139
100	0.01217	0.03622
1000	0.00098	0.01411

Table 1

Improvement over the [X] Subspace Constraint.

For a given desired correlation, the reordered vectors of [X] can be considered to subtend a new [X] subspace and may be used as the starting point for another iteration. Recall [T] transforms the vectors of [X] into a set of vectors, [X'], which have the exact linear correlation prescribed in the desired correlation matrix [S]. Re-ordering the vectors of [X] based on the rank order of the vectors of [X'] is equivalent to selecting the permutations of the vectors of [X] that have the closest possible alignment to the vectors of [X']. That is, the closest possible alignment to a set of vectors whose correlation is exactly [S] ("closest possible alignment" implies maximum inner product of the vector of [X'] and the corresponding re-ordered vector of [X]). Conversely, any permutation of a particular [X] vector that does not lie in the rank order region of the corresponding [X'] vector would be less aligned (lower inner product) with the [X'] vector; resulting in linear correlation values further from the desired correlation.

Consider an input matrix of N X 2. After performing the IC method, the reordered second vector of [X] resides in a rank order region that results in a correlation value close to that prescribed in [S]. With a new [X] subspace defined by the reordered second vector of [X], the IC method is applied once more. Two alternatives may occur:

- The rank order of the resulting second vector of [X'] remains unchanged and so there would be no change to the rank ordering of the second vector of [X]; or,
- Based on the new [X] subspace, the rank order of the second vector of [X'] changes. The first iteration resulted in a good solution. However, since the second vector of [X] no longer has the same rank order of the recalculated second vector of [X'], it is less correlated with the revised exact solution second vector of [X']. Once reordered, it becomes better correlated with the second vector of [X']. As a result, the revised permutation of the second vector in [X] has a resultant linear correlation value even closer to the desired correlation. Any other permutation (particularly the previous one) would be less linearly correlated with the second vector of the revised [X']. Hence, the iteration results in a reduction in resultant linear correlation error.

For NX2 input matrices, iteration until convergence always yields resultant linear correlation values closer to desired with each iteration (when more than one iteration is necessary for convergence). However, it was found that for input matrices of N X M, M>2, improvements with each iteration for each resultant correlation matrix value does not always hold. It will be shown that iterating until convergence reflects improvements for resultant linear correlation; but not necessarily for resultant rank correlation.

The iterative process is described as follows; where the subscript denotes the iteration number:

$[X_1][T_1] = [X_1']$, $[X_1]$ is re-ordered based on the ranks of $[X_1']$ and becomes $[X_2]$.

$[T_2]$ is recalculated from $[X_2]$ and $[S]$.

$[X_2][T_2] = [X_2']$, . $[X_2]$ is re-ordered based on the ranks of $[X_2']$ and becomes $[X_3]$. If there is no change in resultant correlation of the vectors of $[X_3]$, stop. Otherwise…
$[T_i]$ is recalculated based on $[X_i]$ and $[S]$.

$[X_i][T_i] = [X_i']$, $[X_i]$ is re-ordered based on the ranks of $[X_i']$ and becomes $[X_{i+1}]$. Continue iterating until there is no change in resultant correlation of the vectors.

To illustrate this notion, the previous example illustrated in figure 5 serves as the starting point for this iterative process.

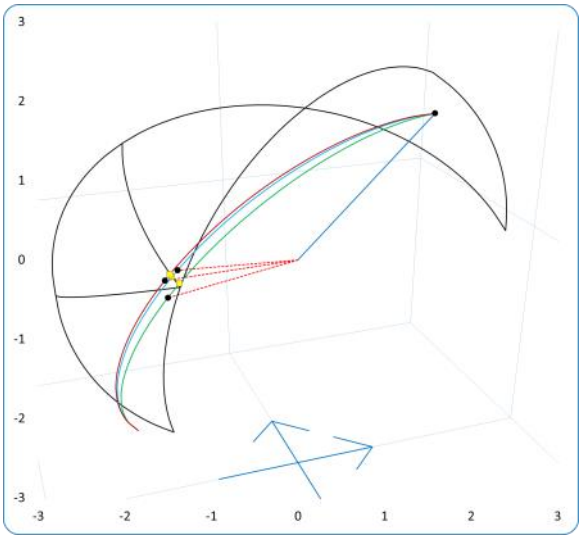

Figure 11

Referring to figure 11 , the blue vector emanating from the center represents the first vector of [X] and its point lies in rank order region 1234. The second vector of $[X_1]$ is the represented as the lower of the 3 red vectors and is mapped to a point in rank order region 3421. The green arc represents the linear combinations of the first and second vectors of $[X_1]$ as determined by [T] through the range of desired correlation values from -1 to 1. The desired correlation remains -.2225.

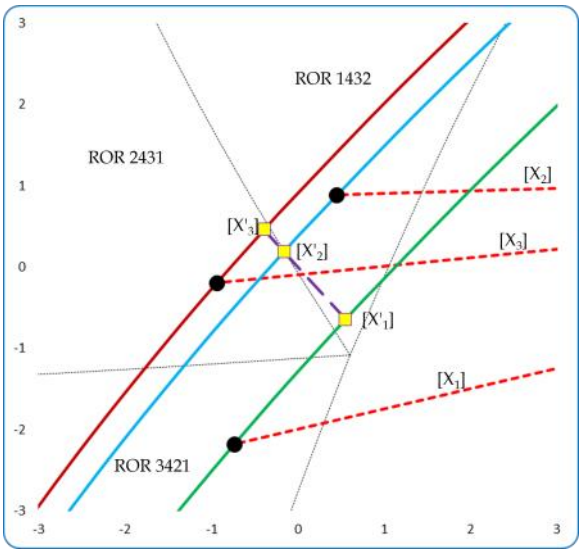

Figure 12

The figure 12 provides a closer view of the area of interest.

- The correlation between the starting vectors of $[X_1]$ is -0.326; yielding a absolute correlation error of 0.103.
- The far right yellow marker represents second vector of $[X'_1]$ whose correlation with the first vector is exactly -0.2225.
- The second vector of $[X'_1]$ lies in a different rank order region (1432). Thus the second vector of [X] is reordered to have the same rank order. This solution is depicted by the upper three red vectors, $[X_2]$. The correlation of the vectors of $[X_2]$ is -0.175; yielding an improved absolute resultant correlation error of 0.048
- Once again, $[X_2]$ is used to calculate $[X'_2]$ where the correlation between the first vector of [X] and $[X'_2]$ is exactly -.2225. In this case, the second vector of $[X'_2]$ has a rank order of 2431; different than that of the second vector of $[X_2]$.
- Thus, the second vector of [X] is reordered to that of the second vector of $[X'_2]$. This is depicted as $[X_3]$. The resultant correlation with this iteration is -.2663; yielding an absolute resultant correlation error of 0.0438; lower than the previous iteration error.

- Another iteration yields [X'$_3$]; whose second vector's rank order remains unchanged.
- Hence the final solution after iteration, [X_3], represents the optimal solution for linear correlation. Importantly, this is predicated on the starting vectors of [X_1] where possible solutions are confined to this subspace. However with iteration, regions beyond the initial [X] subspace may be encountered as was the case in this highly simplified, but wholly representative instance.

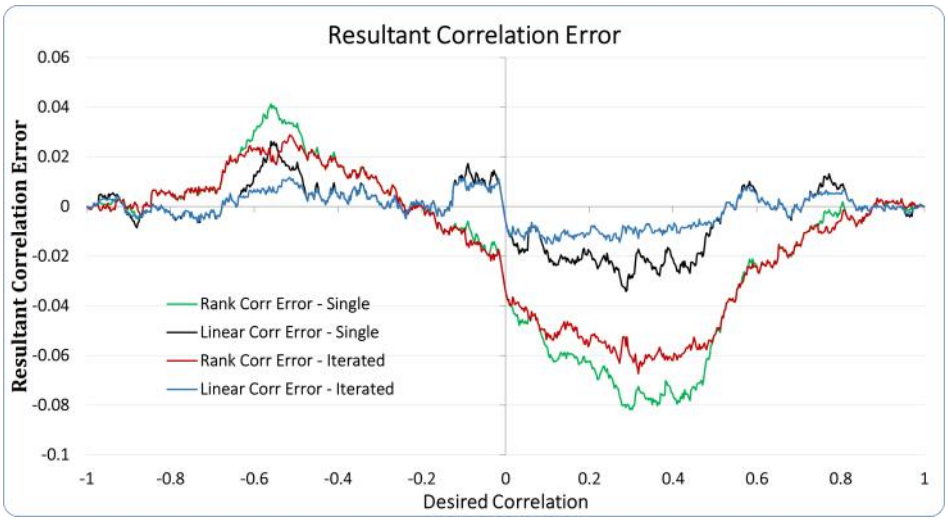

Figure 13: Results for N = 100, Correlation Error Combined

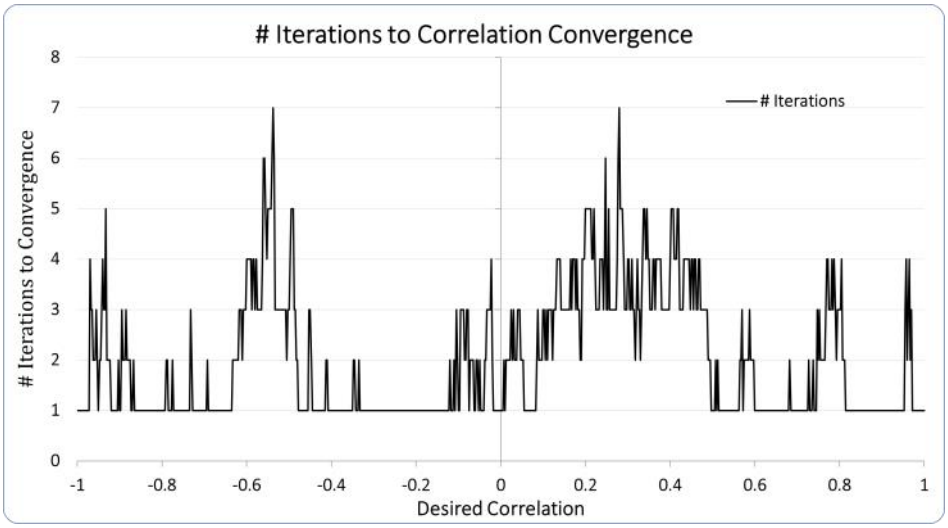

Figure 14: Results for N = 100, Number of Iterations

Correlation with N = 10, 30, 100 and 1000 and Iterating

Similar to before, resultant correlation and resultant correlation error was evaluated for desired correlation values ranging from -1 to 1 in increments of .00125. This now includes results of the iteration process described above for the same four cases of N=10, 30, 100, 1000. Each case used the same starting [X] to equitably compare the single iteration IC method to the iterative process. The following shows the resultant correlation error and number of iterations to convergence for N = 100. The results from the other cases are available as supplemental material for this paper. The chart legend indicates "Single" for applying the IC method once, or "Iterated" which indicates iteration until convergence.

It was found that for all cases where more than one iteration was performed to achieve convergence of the reordered vector, an improvement in linear correlation error resulted. There are, however, ranges of desired correlation where rank correlation error (absolute value implied) worsened (e.g. -.1, .75). The following table shows global error, as previously defined, for all cases and the change in global error from the single IC method to the iterative process discussed here.

	Global Error			
	Single IC Method		Iterating Until Convergence	
N	Linear Correlation	Rank Correlation	Linear Correlation	Rank Correlation
10	0.06457	0.07311	0.05649	0.06464
30	0.03749	0.05139	0.02186	0.03557
100	0.01217	0.03622	0.00651	0.03028
1000	0.00098	0.01411	0.00056	0.01389

Table 2

	Global Error	
	% Change from Single to Iterative Result	
N	Linear Correlation	Rank Correlation
10	-12.52%	-11.59%
30	-41.68%	-30.79%
100	-46.50%	-16.39%
1000	-42.75%	-1.57%

Table 3

The Iteration Method with Skewed Distributions

To stress test the iterative process, three highly skewed input vectors were used as the second vector of [X] for N = 10, 30 and 100. These distributions have a population skewness of 2.62, 3.14 and 8.36; respectively. As before, the first input vector values were derived from the normal distribution. The following graph is a plot of values of the rank ordered second versus the first vector of [X]. Linear correlation of the rank ordered vectors is shown in the legend; quantifying the dissimilarity of the distributions.

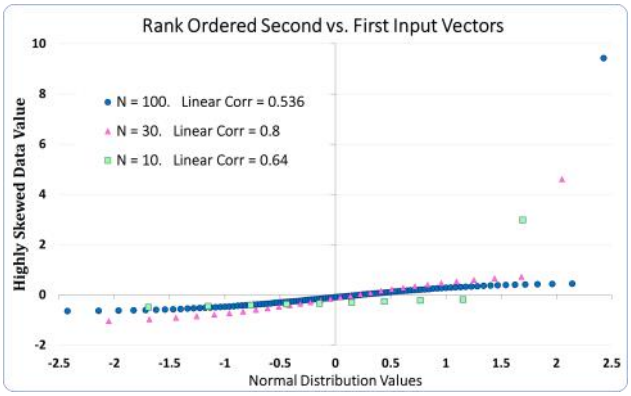

Figure 15

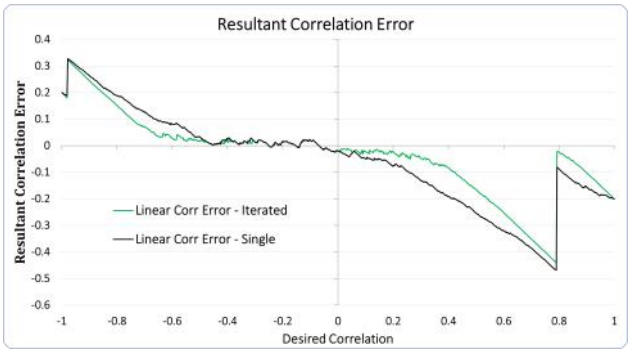

Figure 16: Results for N = 30, Linear Correlation

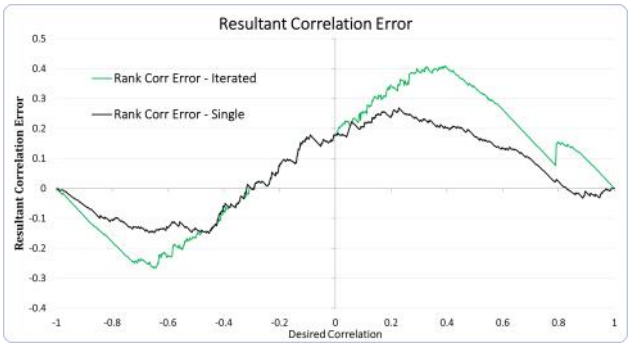

Figure 17: Results for N = 30, Rank Correlation

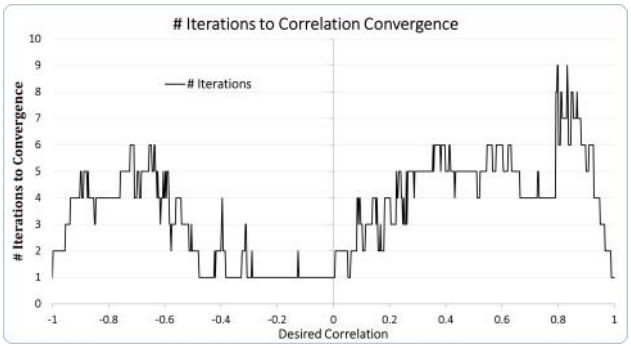

Figure 18: Results for N = 30, Number of Iterations

Results are shown for N = 30. The results from the other cases are available as supplemental online material for this paper.

Observations:

- In all instances, resultant correlation error (absolute value implied) for linear correlation was reduced with more than one iteration until convergence.
- With very few exceptions, resultant correlation error for rank correlation increased with more than one iteration until convergence.
- There is significant error with linear correlation toward the ends of the desired correlation range because of the dissimilarity between the distributions of the vectors of [X].

Higher Dimensional [X] Subspace: Input matrices with 2 variables have been used for illustrating improved linear correlation results of the IC method with iteration. The following considers an input matrix with 16 vectors (i.e. M = 16) for cases of N = 30 and 100. Once again, values are derived from a normal distribution as van der Waerden scores. By virtue of the transform matrix [T] being of upper triangular form, the vectors of [X'] are linear combinations of the corresponding vector of [X] and those to the left. (e.g. the 11th vector of [X'] is a linear combination of the first 11 vectors of [X].) The [X] subspace has expanded to a higher order subspace for input vectors further to the right in the [X]; potentially enabling reduced resultant correlation error. The following steps were taken to assess the effect of the higher dimensional [X] subspace.

- The desired correlation matrix, [S], has all values in the first column and first row the same. That is, all vectors will be seeking the same correlation value with the first vector in order to assess any effect of the higher dimensional [X] subspace. The remaining values of the desired correlation matrix were chosen to ensure positive definiteness.
- Linear correlation error was evaluated with single iteration and with iteration to convergence for desired correlation values ranging from -1 to 1 by .00125. From these results, global error was evaluated for each vector.
- In order to ensure linear independence of the vectors of [X], the 2nd through 16th vectors of [X] were randomly permuted for each desired correlation value.
- Evaluating linear correlation results of the first vector with all others for desired correlation values ranging from -1 to 1 was performed 30 times to attain data with a degree of statistical significance.
- The average and sample standard deviation of the global error from each of the 30 runs, and for each of the 2nd through 16th vectors was evaluated.

The following graphs show the global error results for the 16 vector input matrix [X] with N=30 and N = 100. Also plotted are plus and

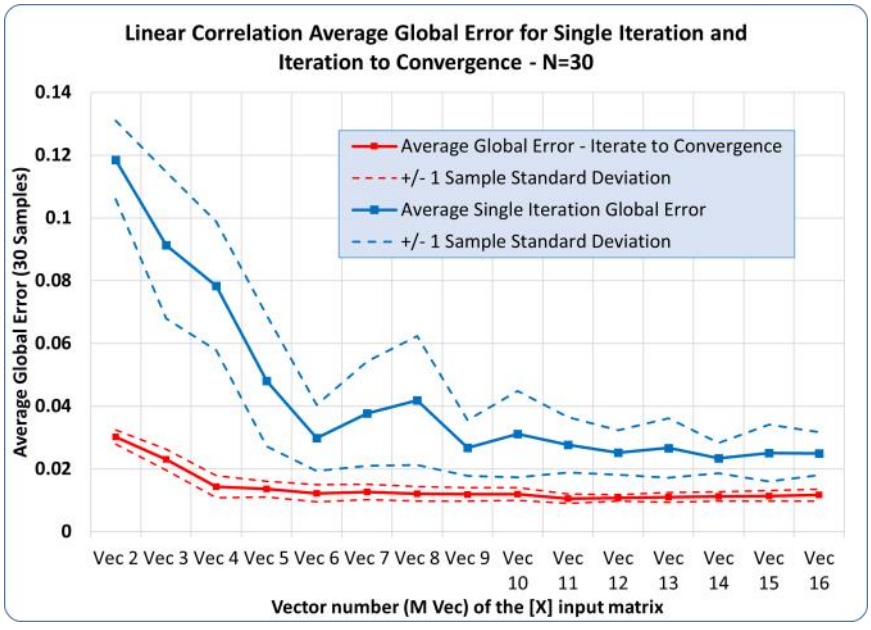

Figure 19

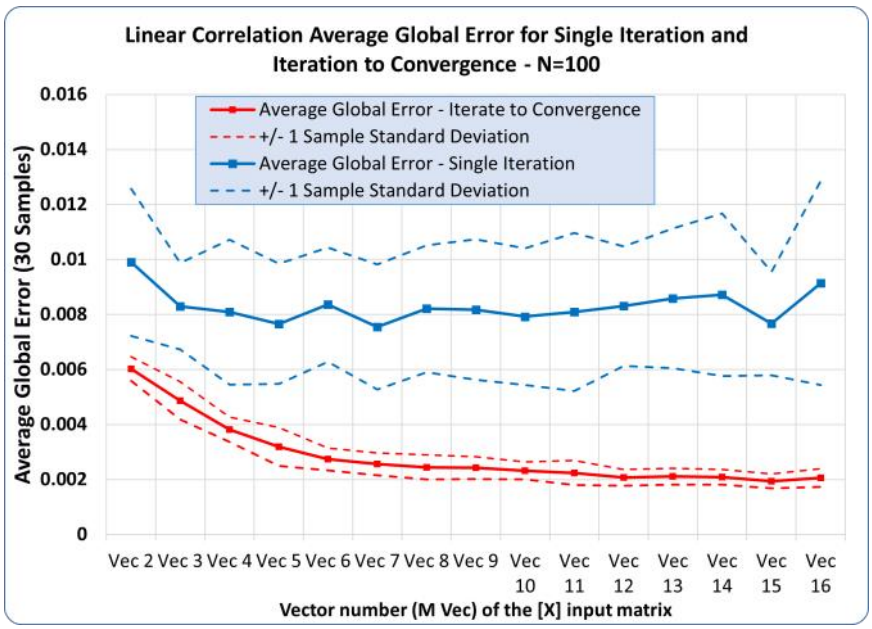

Figure 20

minus one sample standard deviation from the sample mean for the single iteration global error results and for the iteration to convergence global error results based on the 30 runs for desired correlation from -1 to 1 by .00125.

Observations:

- As shown previously in the case of M=2, linear correlation global error is reduced with iteration to convergence.
- For both cases of iteration to convergence for N=30 and N=100, the global error of the 16th vector is approximately one third of that of the 2nd vector. This suggests that iterating within a higher order subspace of N dimensions enables reduced global error. In other words, vectors more to the right in [X] are more likely to have less resultant correlation error than preceding vectors with the iteration method.
- It was observed that while the root mean square of all linear correlation errors reduced with each iteration, some individual correlation errors (i.e. correlation of vector i with vector j; j ≠ i) increased during the iteration process.

Summary

The IC method of developing a multivariate input variable with prescribed correlation is first described and demonstrated. The IC method is then applied for the case of two input variables where the correlation between the two input vectors is prescribed in [S] and the resultant linear and rank correlation of the reordered vectors is calculated. IC method results are evaluated for desired (prescribed) correlation values between -1 and 1 by increments 1/800th. This assessment is performed for input vectors with the number of values (N) ranging from 5 to 1000. A global error is defined as the root mean square of the difference between the desired correlation and that calculated from the IC method for all 801 instances of desired correlation between -1 and 1.

There are N factorial possible permutations of an input vector which span N dimensional space. With all permutations viewed as a constellation of discreet points in N dimensional space, each of these points has a correlation value (linear and rank) with the other input vectors. The IC method constrains possible outcomes to those near the

subspace in N dimensional space subtended by the linear combination of the vectors of [X]. It is possible that there are values in the constellation of possible resultant correlations that are closer to the desired correlation.

An iterative process is presented where the resultant re-ordered vectors of [X] are used as the starting point to once again apply the IC method. This is repeated until convergence of the resultant correlation is achieved. Using the global error previously described as a comprehensive metric, the iterative method shows marked reductions in linear correlation error (i.e. absolute value of the difference between prescribed and resultant correlation) from that of the single step IC method. For highly skewed and dissimilar distributions of input variates in [X], it is shown that rank correlation error often worsens with iteration while linear correlation error improves.

The single iteration Iman Conover method is a powerful technique that is likely more than adequate given the confidence limits of prescribed correlation values. As such the practitioner should consider the value of improving the correlation accuracy of the input variables by iteration.

Having a practical understanding of the methods by which correlated multivariate input variables are developed is useful when the software platform does not provide what is needed. The ability to perform this manually, to understand the nature of its limitations and to experiment with various distribution types may be useful and is certainly instructive for the practitioner. There may be special circumstances where increased accuracy of the correlation of a set of input variables is needed. The author notes that the improvement in accuracy with the application of the iterative IC method described herein should be considered in the context of the confidence limits of the desired correlation coefficient value derived from sampled data where the Fisher r to z transformation has applicability.

References

1. Iman, R. L., Conover, W. J., *A Distribution-Free Approach to inducing Rank Correlation Among Input Variables*. Commun. Statist. Simula. Computa. 11(3) 311-334 (1982)
2. Mildenhall, S., *Correlation and Aggregate Loss Distributions With An Emphasis On The Iman-Conover Method*. Casualty Actuarial Society Forum, Winter 2005. (4) 42-55
3. Tamhane, A. , Dunlop, D. 2000, *Statistics and Data Analysis*, Upper Saddle River (NJ), Prentice Hall Inc. (10) 380-385

Douglas Henke *served in the U. S. Coast Guard from 1980 to 2004. His service included tours as Chief Engineer on the Coast Guard Cutter CAMPBELL, Associate Professor in the Coast Guard Academy's Engineering Department and final tour as Program Manager for recapitalizing the Coast Guard's aging fleet of ships and boats. Since retirement in 2004, he's been a consultant to Program Managers in the Coast Guard's Acquisition Directorate; providing a variety of services including financial and statistical analysis and modeling.*

The International Cost Estimating and Analysis Association is a 501(c)(6) international non-profit organization dedicated to advancing, encouraging, promoting and enhancing the profession of cost estimating and analysis, through the use of parametrics and other data-driven techniques.

www.iceaaonline.com

Submissions:

Prior to writing or sending your manuscripts to us, please reference the JCAP submission guidelines found at

www.iceaaonline.com/publications/jcap-submission

Kindly send your submissions and/or any correspondence to JCAP.Editor@gmail.com

International Cost Estimating & Analysis Association

4115 Annandale Road, Suite 306 | Annandale, VA 22003

703-642-3090 | iceaa@iceaaonline.org