Relationship Between Group Associations and Factor Methods to Correlation

BACKGROUND

This paper addresses assumptions in RI$K regarding the degree of dependence between cost elements. Throughout this paper we will refer back to the following example cost estimating structure.

Appendix Table 1: Example cost Estimating Structure for Study

Cost Element	Random Variable	Distribution Function
Total System Cost		
Total Hardware Cost		
Hardware Item 1	X	f(x)
Hardware Item 2	Y	g(y)
Hardware Item 3	Z	h(z)
Total Support Cost		
Training	T	
SE/PM	S	

Rows such as "Total System Cost" and "Total Hardware Cost" will be referred to as "parent rows" or "aggregate rows." By the indenture structure, it can be seen that these rows are defined as the sum of other rows. Rows such as "Hardware Item #," "Training," and "SE/PM" will be referred to as lowest level items or leaf nodes. In context, they may be referred to as children of their respective parent.

In cost risk analysis, we normally think of each lowest level cost element as a distinct random variable. We specify a target or baseline cost (X, Y, and Z) for each leaf node and a probability distribution (f, g, and h) about each point estimate to characterize the risk. Aggregate level cost elements are then summations of these lower level random variables. Therefore, these aggregate level items are themselves random variables, since a function of one or more random variables is a random variable. The goal of a risk assessment is to estimate the probability distribution of these parent items, given distributions at the lower levels.

In general, the problem of determining the distribution of a function of random variables is a complex, and often unsolvable, problem. Typically, we resort to heuristic methods to approximate these distributions. RI$K uses the Latin Hyper Cube sampling in the Monte Carlo Simulation method, and provides outputs containing mean, standard deviation, and various levels of percentiles. The Latin Hyper Cube method attempts to build up an approximation to the desired distribution from empirical sampling.

An important aspect of dealing with risk analysis is how to treat the risk between or among cost elements that are related to each other (positively, negatively, or both). For example, when the cost of Element X increases, the cost of Element Y should also increase, and Element Z should decrease. This is known as dependency or correlation. Its treatment has varied from models that allow for no associations to those that assume all elements are fully associated. An allied issue is how to treat the relative strength of the relationship among elements. For instance, should the relationship between two elements be either completely associated or should degrees of association be allowed (as in partial correlation)? Roughly, we say that two random variables are independent if a change in one has no effect on the other and vice versa. As has been noted often in the cost risk world, there are many potential sources of uncertainty and risk in a cost estimate. Unanticipated or random events occur which can affect schedules, budgets, technology considerations, etc. When such an event occurs, one or more cost elements are affected. The goal of a cost risk assessment is to capture the likelihood of each possible event, which cost elements are affected, and the magnitude of the impact. Of course, common sense prevails, since it would be impossible to capture all possible events, which can impact cost (e.g., the world coming to an end would have a definite impact on all cost elements, but the likelihood of the event is hopefully very small). In cost risk analysis, asserting that two elements are independent implies that the set of possible events which impact the cost of one have no impact on the cost of the other. Conversely, asserting that two elements are dependent implies that one or more of the same (likely) events impact both cost elements. The assumptions regarding dependence can have a significant impact on the resulting distribution for the parent level cost elements. Mathematically, dependence can be defined in terms of the following theorems.

Theorem 1: If X1,...Xn are random variables and Y=X1 +...+ Xn,

then $VAR(Y) = \sum VAR(Xi) + 2 \sum_{i<j} \sum COV(Xi,Xj)$, where VAR(X) stands for the variance of a random variable, X, and COV(X,Y) is the covariance between two random variables, X and Y.

Theorem 2: If X and Y are independent random variables, then COV(X,Y) = 0.

Theorems 1 and 2 show the importance of the assumptions regarding dependence (or independence). In our example above, Total Hardware Cost is the sum of X, Y, and Z. If these variables are all dependent (and so $COV(X,Y) \neq 0$, etc.), but we mistakenly ignore this fact, then we incorrectly estimate the variance at the aggregate level as VAR(X) + VAR(Y) + VAR(Z). This underestimates the true variance by 2(COV(X,Y) + COV(X,Z) + COV(Y,Z)). In this case, the true distribution would be broader than our estimate. Note, however, that asserting that X and T are dependent will only affect the resulting distribution at the Total System Cost level. Total Hardware Cost and Total Support Cost will not be affected. Furthermore, dependency assumptions have no effect on the expected value of the distribution.

In practice, it is unlikely that a total weapon system cost estimate will consist of either completely interdependent or independent cost elements. However, it should be noted that the desired solution (partial interdependence) is bounded by the solutions to these extreme cases. By exploring these two cases, the analyst can determine the relative importance of these assumptions. Complete interdependence would provide an upper bound on the desired

solution (or we could say a pessimistic estimate). In most cases, these extreme solutions will be unsatisfactory. The low estimate (independence) will be too low, and the high estimate (dependence) will be too high. The analyst should typically attempt to determine where the true estimate falls.

This mathematical definition of dependence strongly suggests that the notion of dependence in cost risk analysis can be treated as a problem in correlation with the "degree" of dependence measured as a correlation coefficient. Although this may be a practical construct from which to view the problem, there is extreme difficulty in arriving at appropriate correlation coefficients. While it is true that correlations between cost elements can easily be calculated, this correlation between cost elements is not appropriate for cost risk assessment. In risk analysis, when we ask if two cost elements "move together," we are actually asking if there is some common factor that causes our uncertainties in the estimates of these individual elements to move together, not the estimates themselves. This uncertainty is reflected in the residuals or percentage errors of our estimating process, so, in essence, we are asking if these residuals or percentage errors are correlated. Strong correlation between cost elements in a database should not be mistaken as evidence that the noise terms of our estimating process, possibly derived from this same database, are correlated. Finding correlation in a cost database is the job of building CERs and of the estimating process. The noises about these CERs are assumed to be random if the correct driver variables are used.

The dependencies of uncertainties in a cost risk analysis arise because of the way in which a program or project is structured. If two or more activities are scheduled to occur concurrently when they would ordinarily be accomplished in series, then problems in one task may affect the other. Similarly, activities that are all affected by a common technology shortfall may exhibit common cost impacts associated with redressing the shortfall. The dependence between the uncertainties of estimates for elements of the WBS is determined by the structure of the development process. This dependence is not intrinsic to the elements themselves.

These characteristics that give rise to correlations among uncertainties, schedule concurrency, and common technology difficulties, are assumed to be unique for each program or project. In a cost database of comparable programs, these factors may have significant cost impact on some programs and little impact on others. A CER built from this database captures the mean effect of all these factors over all programs. These program characteristics that give rise to uncertainty add to the variance of any CERs developed. Since these characteristics are unique to each program, there is no reason to believe that they will be systematic across all data points in the database; some factors will have a significant impact on some programs and not on others. To try to discover the interdependence of our uncertainties in the estimates of cost elements from a historical database, we would need to first isolate or "normalize out" those factors with cause uncertainty from each data point. Such normalization would be, at best, tedious, time consuming, and heavily reliant on a subjective evaluation of each data point.

Given that correlations (of uncertainties) will be difficult, if not impossible, to develop objectively from databases, how does one go about it for the purpose of risk modeling? Since the characteristics giving rise to interdependence among uncertainties are unique to each program, the analyst must make a determination of which elements are likely to move together and by what degree. These are subjective determinations that will be guided by the

analyst's investigation of the program structure and potential technology problems. Paul Garvey of MITRE suggests[2] a subjective set of criteria for assigning correlations (with some modification for inverse or negative relationships):

Appendix Table 2: Subjective Set of Criteria for Assigning Correlations

Description	Correlation Assignment
No interdependence between elements	0
Some interdependence between elements	0.25
Moderate interdependence between elements	0.5
Strong interdependence between elements	0.75
Complete interdependence between elements	1

This subjective approach allows the analyst to make a judgment if the uncertainties about two elements should move together, in the same or opposite directions. If a relationship exists, subjective judgment is used to set the strength. The analyst needs to consider all of the programmatic, technical, schedule, and budgetary factors to make these judgments. Note again that, while correlation provides a convenient context for discussing dependencies, we are not suggesting that correlation coefficients between cost elements, derived from a cost database, should be used. Analyst judgment is needed.

The remainder of this paper presents a new heuristic approach, called "group association," for handling interdependence between cost elements. This approach has been included in the Tecolote Research cost risk model, RI$K. Section 2.0 provides an algorithmic description of the method. The group association heuristic is discussed in the context of the simulation model. Empirical results are presented in section 3.0. We attempt to show the impact of groupings on the resulting convolved distributions, given a number of different starting assumptions. We also present the resulting interterm correlations between related elements. Conclusions are discussed in section 4.0, together with some suggestion for future research.

ALGORITHMIC DESCRIPTION OF GROUPING HEURISTIC

Loosely, we define a *Group* as a collection of cost elements that are all pairwise related. Both positive (items move in the same direction) and negative (opposite direction) relationships are allowed in the same group. In our above example, elements X and S might be positively related to one another and negatively related to Y (say due to various schedule and technology requirements). In this case, X and S might both have relatively high costs while Y would have a relatively low cost (all in unison), or X and S might be low while Y is high. Note then that elements are grouped together if there is a defined relationship (schedule, technology, programmatic...) between them such that events (problems or lack of problems) that affect the cost of one will (most likely) affect the costs of the others. The notion of association strength defines the degree to which this relationship is deterministic. For each element of a group, an associated strength is assigned. As with correlation, strength must be between -1.0 and 1.0. The sign indicates the relative direction of the relationship, and the unsigned value measures the magnitude of the impact. The exact magnitude is determined by the convolution method used.

SIMULATION MODEL

As we have discussed, the impact of dependencies and groupings occurs during the convolution process. When building to a parent level distribution from lower level cost elements, relationships between the lower level items can have a direct impact on the resulting distribution. A common approach to estimating the convolved distribution is the Monte Carlo simulation method.

Ignoring grouping, a Monte Carlo approach would typically make a random draw, independent of all other draws, for each lowest level cost element. In RI$K, these draws correspond to confidence levels. For each leaf node, we select a value between 0% and 100%. This draw value is then converted into the cost of achieving the corresponding level of confidence. So if the draw for an element is 95%, then the result would be the cost necessary to complete the task with 95% confidence. In our example, cost values are calculated from draws for variables X, Y, Z, S, and T. The cost values for Total Hardware, Total Support, and Total System are then calculated as the appropriate sums (according to indenture). This process is repeated many times to develop an approximation of each aggregate element's distribution.

With grouping, the process is somewhat more complicated. For each Monte Carlo iteration, RI$K first sets a target confidence level for each group. Group targets are also random values, independent of each other, between 0% and 100%. The group target defines the starting point for determining the draws for each element of the group. Each group element's strength value defines how "close" the element's draw is to the target draw. The element's draw is randomly selected in the range defined by the target draw ± 100*(1.0 - Strength) (with a little bookkeeping for boundary conditions and negative strength values). So a strength of 1.0 would force the element to have the same draw value as the target, while a strength of 0.0 would put no limitation on the element's draw. As before, the element draws are then converted into costs that are summed to aggregate levels.

While this approach seems deceptively similar to the notion of correlation, it should not be thought of as such. With the grouping approach, it is important to realize that association strength is a measure of the closeness of the random draws or *confidence levels*. With the above-described procedure, two positively related elements will both have costs at approximately the same level of confidence; they move together. For some Monte Carlo iterations, both costs might be at the 5% level, while on others they might be at the 75% or 95% levels. From this, we can conclude that the respective costs will be high or low together. But this is not necessarily a linear relationship. The form of the relationship is determined not only by the strength value but also by the specified distributions for each item. If one was specified as Uniform and the other Normal and the strength was set to 1.0, the relationship would certainly not be linear (but essentially Normal). Empirical results in section 3.0 illustrate these ideas.

Two additional concepts should be discussed. First, RI$K allows one element of each group to be distinguished as the dominant element. We think of this element as the key cost element or driver. Random events that affect this element cause chain reactions to the other elements in the group. In this situation, the dominant element's draw becomes the target draw for the other elements in the group. The second idea that needs to be discussed is that of a factor method. RI$K allows leaf nodes to be identified as simple factors of other rows. In our example, we might define S = a*X and T = b*(Total Hardware Cost). After a cost is

calculated for X (based on the confidence draw for X), the cost for S is calculated by applying the factor relationship. The cost for T is calculated using the factor relationship applied to the summed cost for Total Hardware. Since these are direct linear relationships, it is important to note that these factor rows, S and T, are correlated with the drivers, X, Y, and Z, and to each other (even if the exact correlation coefficient is not specified or known). This correlation will then have an impact on the aggregate level convolutions, Total Support Cost and Total System Cost. Similarly, when the input physical parameters, i.e., weight, power, etc. are specified by probability distributions, the configuration risk can be handled stochastically.

EMPIRICAL RESULTS

Three distinct sets of experiments were performed. The first two sets illustrate the impact of group associations on the convolution process. Interterm correlations are also calculated. The first set of four cases all use the same RI$K inputs, except for group strength assignments. The second set of four cases are identical to the first set, except that skew assumptions have been varied. Results are presented using 2000 iterations.

EXPERIMENTS 1 THROUGH 4

RI$K Inputs

Appendix Table 3: Distribution Inputs Experiments 1-4

Distribution Inputs (Exp. 1 - 4)					
Row	**WBS**	**I/P Cost**	**Distrib.**	**Skew**	**Spread**
1	Total				
2	Part 1	100	TRIANG	RIGHT	HIGH
3	Part 2	100	BETA	RIGHT	HIGH
4	Part 3	100	TRIANG	RIGHT	HIGH
5	Part 4	100	UNIFORM	RIGHT	HIGH

Appendix Table 4: Association Strengths by Experiment

Association Strengths by Experiment (All Items in Same Group)					
Row	**WBS**	**Exp. 1**	**Exp. 2**	**Exp. 3**	**Exp. 4**
1	Total				
2	Part 1	1	1	1	0.8
3	Part 2	1	0.9	1	0.8
4	Part 3	1	0.8	-1	0.8
5	Part 4	1	0.7	-1	0.8

Results

Appendix Table 5: Confidence Level Results for Total

Confidence Level Results for Total Row (Row 1)						
Case	**Mean**	**50%**	**70%**	**90%**	**95%**	**99%**
w/o Grouping	508.4	507.8	545.2	602.3	626.7	681.8
Monte Carlo						
Exp. 1	508.4	494.6	589.3	714.7	761.6	822.2
Exp. 2	508	494.7	593.7	696.3	732.8	779.8
Exp. 3	508.4	504.1	513.9	527.6	534	555.4
Exp. 4	507.2	497.8	590.7	691	715.6	754

Appendix Table 6: Inter-Item Correlations for Experiment 1

Inter-Item Correlations for Exp. 1				
	Part1	**Part2**	**Part3**	**Part4**
Part1	1			
Part2	0.999	1		
Part3	0.9991	0.999	1	
Part4	0.985	0.989	0.985	1

Appendix Table 7: Inter-Item Correlations for Experiment 2

Inter-Item Correlations for Exp. 2				
	Part1	**Part2**	**Part3**	**Part4**
Part1	1			
Part2	0.968	1		
Part3	0.894	0.877	1	
Part4	0.808	0.798	0.752	1

Appendix Table 8: Inter-Item Correlations for Experiment 3

Inter-Item Correlations for Exp. 3				
	Part1	**Part2**	**Part3**	**Part4**
Part1	1			
Part2	0.999	1		
Part3	-0.97	-0.97	1	
Part4	-0.99	-0.99	0.985	1

Appendix Table 9: Inter-Item Correlations for Experiment 4

Inter-Item Correlations for Exp. 4				
	Part1	**Part2**	**Part3**	**Part4**
Part1	1			
Part2	0.825	1		
Part3	0.818	0.821	1	
Part4	0.835	0.837	0.838	1

Observations

As expected, the mean estimates for these cases are all approximately the same as the estimate without groupings. With groupings, if all elements move together (Exp. 1, 2, and 4), then the parent distributions are spread out (greater risk). The cost of achieving a 99% level of confidence varies from 681.8 (without groupings) to 822.2 with complete dependence (Exp. 1). This gives a potential increase of 20.6%. When two of the elements in the group have a negative strength, as expected, the parent distribution "shrinks" (99% point for Exp. 3 is 555.4). This gives a potential decrease of 18.5%.

Finally, from the correlation tables, we can see that inter-item correlations also track well to the user-specified group strength values.

EXPERIMENTS 5 THROUGH 8

RI$K Inputs

Appendix Table 10: Distribution Inputs Experiments 5-8

Distribution Inputs (Exp. 5 - 8)					
Row	**WBS**	**I/P Cost**	**Distrib**	**Skew**	**Spread**
1	Total				
2	Part 1	100	TRIANG	RIGHT	HIGH
3	Part 2	100	BETA	RIGHT	HIGH
4	Part 3	100	TRIANG	LEFT	HIGH
5	Part 4	100	UNIFORM	LEFT	HIGH

Appendix Table 11: Association Strengths by Experiment

Association Strengths by Experiment (All Items in Same Group)					
Row	**WBS**	**Exp. 5**	**Exp. 6**	**Exp. 7**	**Exp. 8**
1	Total				
2	Part 1	1	1	0.9	0.8
3	Part 2	1	1	0.9	0.8
4	Part 3	1	-1	0.9	0.8
5	Part 4	1	-1	0.9	0.8

Results

Appendix Table 12: Confidence Level Results for Total

Confidence Level Results for Total Row (Row 1)						
Case	**Mean**	**50%**	**70%**	**90%**	**95%**	**99%**
w/o Grouping	393.7	394.8	432.3	488.9	519.7	559.7
Monte Carlo						
Exp. 5	391	386.8	477	588.2	629.3	682.2
Exp. 6	391	389.6	392.2	393.8	401.2	429
Exp. 7	391.1	384.8	478.3	587.4	616	646
Exp. 8	391.3	390.5	480.5	565.9	588.2	622.2

Appendix Table 13: Inter-Item Correlations for Experiment 5

Inter-Item Correlations for Exp. 5				
	Part1	**Part2**	**Part3**	**Part4**
Part1	1			
Part2	0.999	1		
Part3	0.97	0.974	1	
Part4	0.985	0.988	0.99	1

Appendix Table 14: Inter-Item Correlations for Experiment 6

Inter-Item Correlations for Exp. 6				
	Part1	**Part2**	**Part3**	**Part4**
Part1	1			
Part2	0.999	1		
Part3	-0.998	-0.997	1	
Part4	-0.985	-0.988	0.99	1

Appendix Table 15: Inter-Item Correlations for Experiment 7

Inter-Item Correlations for Exp. 7				
	Part1	**Part2**	**Part3**	**Part4**
Part1	1			
Part2	0.946	1		
Part3	0.927	0.932	1	
Part4	0.944	0.948	0.952	1

Appendix Table 16: Inter-Item Correlations for Experiment 8

Inter-Item Correlations for Exp. 8				
	Part1	**Part2**	**Part3**	**Part4**
Part1	1			
Part2	0.825	1		
Part3	0.821	0.827	1	
Part4	0.835	0.841	0.845	1

Observations

The results here are also as expected. Note that the mixed SKEW assumptions (some group elements skewed LEFT and some skewed RIGHT) do affect the results somewhat, but the results are still consistent. If group elements all move together, there is a potential cost increase of 21.9% at the 99% confidence level. If two elements move against the group (Exp. 6), there is a potential cost decrease of 23.4%. Correlation coefficients again track quite closely to strength values.

THIS PAGE IS INTENTIONALLY LEFT BLANK