Correlation Matrices Revisited

Steven Ikeler
Army Capabilities Integration Center
2013 ICEAA Conference, New Orleans

Agenda

- Introduction
- Definitions and Uses
- Examples and Properties
- Recent Results and Implications
- C-BA Sensitivity Analysis
- Scenario Analysis Technique
- Follow on Topics and Questions

Introduction

When combining several risks in the model, it's important to account for correlation. Most models depend on user defined risk variables and user input correlations. These user inputs can have a significant impact on risk adjusted estimates.

Introduction

Correlation Matrices are a frequent SCEA/IPSA/ICEAA topic:

- Pairwise Correlation Matrices are using for simulating correlated random variables
- There's a lot of information in the matrix
- Used in estimates with cost risk
- This presentation will discuss techniques for credibly completing a correlation matrix

Definitions and Uses

Correlation is important in the estimate:

- Risks are usually modeled as separate distributions
- The sum of n independent risks each with variance v is much less than the sum of the same n risks with positive correlation
- Be careful adding risk detail. Done incorrectly it arbitrarily increases total cost risk
- Reference: CEBok module 9

Definitions and Uses

n correlated distributions can be simulated by Monte Carlo simulation:

- Given n distributions v_1 to v_n and pairwise correlation matrix M, m_{ij} is the correlation between distributions i and j
- Generate n independent samplings $u_1(0,1)$ through $u_n(0,1)$
- Calculate $w = Ru$ where $R^T R = C$ is the correlation matrix (R is lower triangular)
- Use w to generate each of the v_i using inverse cdf
- Inconsistent matrices may not even have the necessary decomposition

Definitions and Uses

There are other uses for correlated inputs besides Cost Risk:

- Risk Portfolios
- Tradespace analysis with multiple components
- Risk allocation within a cost estimate
- Allocating risk and management reserves
- Modeling dependencies
- Cost risk is more than cost uncertainty

Example

When combining two risk adjusted estimates, we commonly encounter the situation of how to combine the correlation matrices

$$R_1 = \begin{bmatrix} 1 & .3 \\ .3 & 1 \end{bmatrix} \quad \text{and} \quad R_2 = \begin{bmatrix} 1 & -.1 \\ -.1 & 1 \end{bmatrix}$$

R_1 and R_2 are two correlation matrices for two estimates each with two primary risks. They combine as:

$$R = \begin{bmatrix} 1 & .3 & & \\ .3 & 1 & & \\ & & 1 & -.1 \\ & & -.1 & 1 \end{bmatrix}$$

Properties

- In the previous example, filling the missing information with 0's is valid
- A more common problem is validating user-input data
- Correlation matrices are positive semi-definite (eigenvalues are non-negative)
- If the matrix M is a correlation matrix,

$x^tMx > 0$ for all vectors x

- Diagonal of a correlation matrix is 1, all others must be between 1 and -1

Properties

- The convex combination of 2 nxn correlation matrices is again a correlation matrix
- The correlation between two variables is completely independent of the other variables

Trivial Examples:

- Identity Matrix has all eigenvalues of 1
- The matrix with all non-diagonal entries of 1 has eigenvalues of 0 and 1
- If variables 1 and 2 have positive correlation and 2 and 3 have positive correlation, do 1 and 3?

Recent Results

- Heuristic Correction Techniques (uncertainty whether the user-entered data is consistent)
- A search found over 100 results on completing Positive Semidefinite and/or Correlation Matrices in the last 10 years
- Most derive some conditions on the connectedness of the data and the entries that are known and assume they are correct
- NP Hard

Implications of Recent Results

- If you use correlation in a estimate with third-party software, there's a lot of documentation to read
- Completing a correlation matrix is typically a problem for the software designer to validate user-input data, but could be a problem for the estimator to model correctly to begin with
- User input data still lacks credibility

Recent Results (Techniques)

Typical risk correlation techniques use historic programs:

- DoD Selected Acquisition Reports tend to have the information in them (high level risk)
- May not get to the key issues that drove risk
- These use the effects to guess what the causes were
- Current programs are more relevant to future programs than 10 year old programs
- Doesn’t consider programs that were cancelled

Recent Results (Techniques)

Other methods use SMEs:

- SMEs good at giving high, most likely and low estimates for a variable, not necessarily experts at estimating correlation
- SMEs usually don't have the historical knowledge of several historic programs
- Most cost estimators have trouble explaining correlation between two continuous variables
- SME methods are not repeatable
- SMEs are comfortable estimating the effects of programmatic decisions

C-BA Decision Matrices

In the C-BA Presentation, we developed cost per capability:

- A natural question was to look at the correlation between benefits during the sensitivity analysis
- If risk areas replace “benefits”, the results are meaningful for risk correlation
- Some information on the individual risk distributions can be found, as well

C-BA Decision Matrix

- In the example yesterday, we are really talking about performance trades that can be made
- A cost estimate was generated for each alternative
- Recommendation based on Cost Benefit Index

	Wt	Alt 1	Alt 1 Wtd Score	Alt 2	Alt 2 Wtd score	Alt 3	Alt 3 Wtd Score	Alt 4	Alt 4 Wtd Score
Range	.5	1	**.5**	5	**2.5**	5	**2.5**	5	**2.5**
Payload	.25	3	**.75**	4	**1**	4	**1**	5	**1.25**
Weight	.25	2	**.5**	5	**1.25**	3	**.75**	5	**1.25**
Total			**1.75**		**4.75**		**4.25**		**5**

C-BA decision matrix technique

- If the benefits are replaced with risk variables, then an estimated correlation between risks can be found

	Wt	Alt 1	Alt 1 Wtd Score	Alt 2	Alt 2 Wtd score	Alt 3	Alt 3 Wtd Score	Alt 4	Alt 4 Wtd Score
Range	.5	1	**.5**	5	**2.5**	5	**2.5**	5	**2.5**
Payload	.25	3	**.75**	4	**1**	4	**1**	5	**1.25**
Weight	.25	2	**.5**	5	**1.25**	3	**.75**	5	**1.25**
Total			**1.75**		**4.75**		**4.25**		**5**
Cost			**300**		**370**		**340**		**400**

Scenario Analysis Technique

The Risk Modeling Technique is:

- A natural follow-up to the C-BA decision matrix
- Begins with the estimates from the C-BA
- Further refine the alternatives and add the desired risk categories
- Consider different acquisition strategies and risk management strategies
- Develops most likely cost, performance and schedule outcome in each case

Scenario Analysis Technique

Step 1: Make a list of the risks (variables) you are modeling and decompose into independent sets and combine into meaningful summary variables

Step 2: Develop alternatives (scenarios) within the program tradespace that induce variations among those variables

Step 3: Develop estimates for each of the alternatives (find variable values)

Step 4: Validate the observation set and use weights if desired

Step 5: Calculate the correlation between variables (the scenarios are observations)

Scenario Analysis Technique

Types of risk that can be modeled (government perspective):

- Schedule risk (directly)
- Configuration risk (trades within the allowable end product)
- Acquisition risk (use different Acquisition strategies)
- Affordability risk (probability of cancellation)
- Technical risk

Scenario Analysis Example

- The columns are observations, and we can find the correlation between columns (variables)
- The variables are objects (lines) in the cost estimate that use risk distributions

	Alt 1	Alt 2	Alt 3	Alt 4	Alt 5	Alt 6	Alt 7	Alt 8
Annual O&S Cost ($K)	30	15	10	7.5	20	10	5.8	4.7
Unit Cost ($K)	350	350	350	299	278	278	266	234
Schedule	13	17	20	20	12	16	19	19

Scenario Analysis Example

In this example, the scenarios were built on different mixes of acquisition strategies and contract incentives, the variables are the results of those mixes.

	Alt 1	Alt 2	Alt 3	Alt 4	Alt 5	Alt 6	Alt 7	Alt 8
Annual O&S Cost ($K)	30	15	10	7.5	20	10	5.8	4.7
Unit Cost ($K)	350	350	350	299	278	278	266	234
Schedule	13	17	20	20	12	16	19	19

A distribution finder should be run on the variables to ensure the appropriate number of observations.

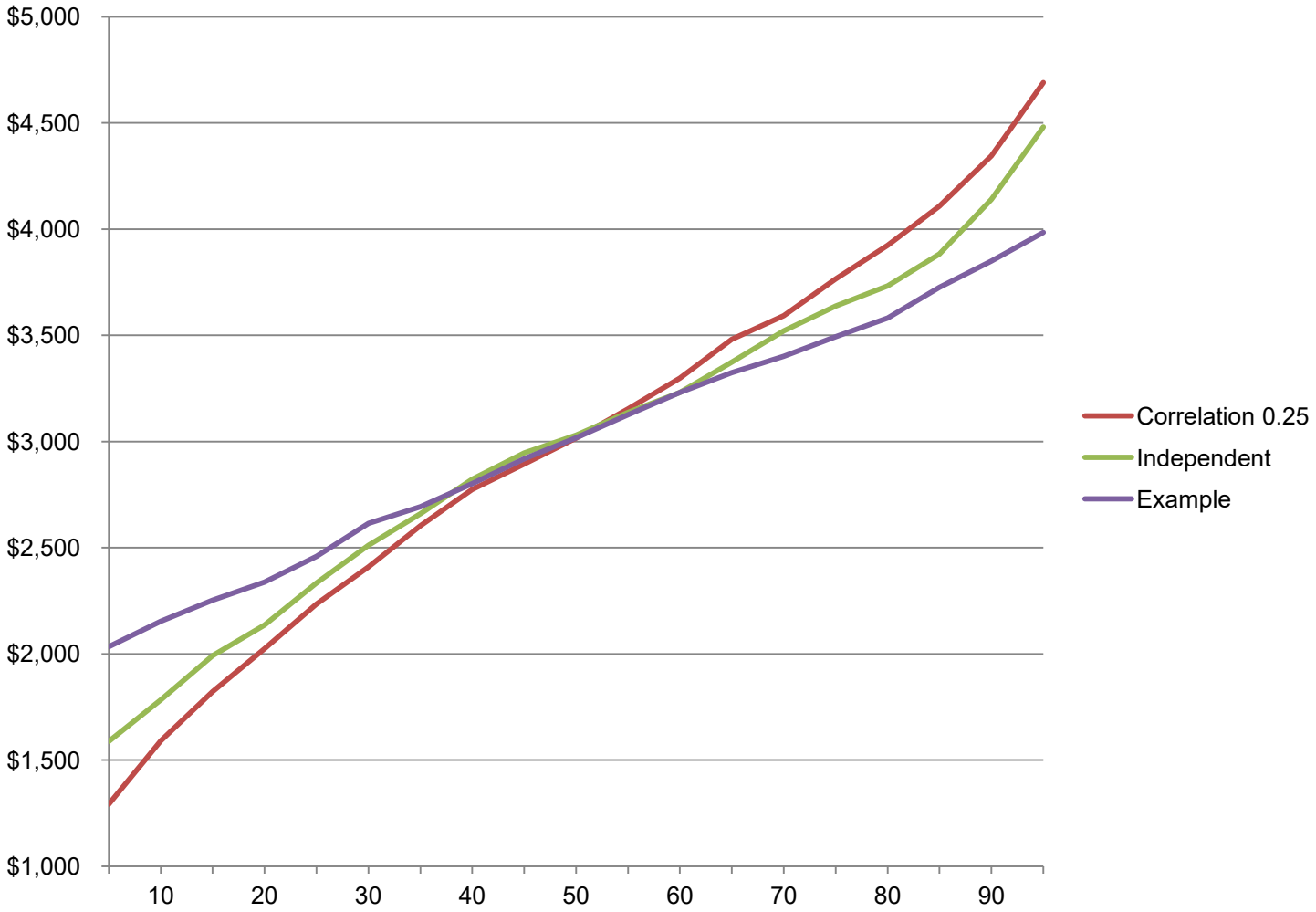
Scenario Analysis Example

In the simple example, the correlation matrix is:

	Annual O&S Cost	Unit Cost	Schedule
Annual O&S Cost	1	0.5660	-0.8320
Unit Cost	0.5660	1	-0.0899
Schedule	-0.8320	-0.0899	1

Scenario Analysis Example

S-curve from the example applied to the simple sum of three uniform distributions. Compare with uncorrelated data and 0.25 correlated.



Scenario Analysis Example 2

In this example, the scenarios were built on different product configurations that meet the contract requirements and/or user utility

	Alt 1	Alt 2	Alt 3	Alt 4	Alt 5	Alt 6	Alt 7	Alt 8	Alt 9	Alt 10
Drivetrain	89,200	81,700	73,100	108,400	104,900	97,500	122,300	118,500	98,700	165,300
Armor	186,300	141,100	55,800	136,200	167,300	118,500	183,000	140,000	109,100	245,300
Suspension	159,500	139,100	131,800	159,100	132,800	126,100	159,400	136,000	128,200	162,300

A distribution finder should be run on the variables to ensure the appropriate number of observations.

Correlation Matrix	Engine	Transmission	Suspension
Drivetrain	1	0.7774	0.5260
Armor	0.7774	1	0.7156
Suspension	0.5260	0.7156	1

Scenario Analysis Challenges

Its difficult to overcome the temptation to overdo the number of observations:

- When the rows exceed the number of columns, our statistics package causes an error message for rank during regression calculations
- Only two rows are necessary for determining the correlation between them
- More observations is better. Want the variable values span the modeled variable

Scenario Analysis Technique

The benefits of this technique:

- Each alternative and estimate is explainable (defensible)
- The scenario strategies can be costed
- Provides information to the Program Manager
- The SMEs have the data to differentiate between true uncertainty and programmatic risk
- The resulting cost distribution can be used to validate the overall cost risk confidence level
- SMEs are good at consistency among alternatives

Scenario Analysis Technique

The difficulties of this technique:

- The alternatives require a lot of documentation
- All the risks that you want correlation should be included in the simulation and its tempting to assume the number of observations needs to exceed the number of risks
- The alternatives should be weighted based on their likelihood
- Should be rooted in real examples
- Don’t try to model true uncertainty

Scenario Analysis Results

The resulting correlation matrix may be used directly:

- It is based on non-continuous variables
- Can also be used to “fill in” missing or inconsistent correlation data
- I claim the results should be used directly

Interpreting the Results

To complete a partial data correlation matrix M' with our credible correlation matrix C:

- $C=LL^t$ where L is the lower triangular matrix from the Cholesky decomposition.
- Given completed M' correlation matrix, there is a lower triangular C' so $M'=C'C'.^t$
- For any lower triangular C', $C'=L'L$ for some lower triangular L'
- $M' = C' C' = L'L\ L^t L'^t$
- Therefore, we are guaranteed there is a way to complete M based on C

Follow-on Topics (Risk Modeling)

- DoD has well defined risk scales (1-5) and experience
- These can be used for risk correlations

	Alt 1	Alt 2	Alt 3	Alt 4	Alt 5	Alt 6	Alt 7	Alt 8
Schedule	1	1	1	2	2	2	1	3
Unit Cost	3	3	2	3	4	3	2	4
Weight	4	4	2	5	3	4	5	3

Follow-on Topics

Ordinary linear regression can be used, with complete cost estimates, to develop top-level cost per risk estimates

	Alt 1	Alt 2	Alt 3	Alt 4	Alt 5	Alt 6	Alt 7	Alt 8
Performance	1	1	1	2	2	2	1	3
Schedule	3	3	2	3	4	3	2	4
Unit Cost Growth	4	4	2	5	3	4	5	3
Cost	**200**	**250**	**300**	**270**	**260**	**250**	**200**	**300**

Follow-on Topics

Cost per risk methodology:

- Assumes cost risk is comparable with cost of avoiding it
- Up-front modeling challenge
- Cost estimate should be modeled to mitigation activities and reserves

Questions