Uncertainty in Traffic Crash Reconstruction, Part 2: Dealing with Correlation

Jeremy S. Daily

June 10, 2008

The results of a traffic crash reconstruction often include vehicle speeds to address causation and changes in velocity to indicate vehicle severity. Since these results are related, they must be modeled in a probabilistic context as a joint distribution. Current techniques in the traffic crash reconstruction literature, however, ignore the correlation structure associated with both the input and results of an analysis. Therefore a discussion of the uncertainty propagation techniques with correlation and Monte Carlo simulation of correlated variables is presented. The idea that measuring a parameter with a common instrument induces correlation is explored by examining the process of determining vehicle weights. Also, an example of determining the energy from crush is presented since the $A$ and $B$ stiffness coefficients are highly correlated. Results show the difference between accounting for correlation and assuming independence can be significant. Furthermore, interpreting and presenting results from simple Monte Carlo analysis of a momentum problem requires using the concepts of joint, marginal, and conditional distributions to fully understand the results.

1 Introduction

Events leading to a traffic crash may become the question of legal action and must be reconstructed by an expert. However, the expert usually does not have perfect information or modeling procedures to determine an exact result. As such, every analysis of a crash must be done knowing that the final result may not be the actual answer. Since these analysis results are not exact, experts must qualify their answer by providing a range of likely values. To this end, many authors have presented techniques for dealing with uncertainty [1–7].
The range of probable analysis results, whether mathematically determined or not, stems from some sort of an uncertainty analysis. Classic uncertainty analysis uses derivative information to map the input uncertainty range to an output range. The other popular technique in the crash reconstruction literature is the Monte Carlo method [8–14]. However, the applications described in the literature rarely describe the techniques or consequences of correlation. Therefore, the purpose of this paper is to provide the analyst with some tools for performing an uncertainty analysis and interpreting the results of an analysis involving correlated variables.

It is very important to realize that there is no mathematical technique available to help the reconstruction expert recover from a blunder. In other words, throughout this paper, it is assumed that the analysis performed was applicable and done correctly. Good reconstruction practice uses multiple techniques to determine values of interest in a reconstruction, if possible. Also, it is assumed that the reader is familiar with crash reconstruction techniques and has some training in the field. While the examples of this paper uses simple equations that are tractable by hand, computer based reconstruction programs can also benefit from the concept dealing with correlation in an uncertainty analysis.

In Part 1 of this series, the generation of an input distribution from small numbers of data was presented. The concepts in this part of the paper use those distributions as inputs for the analysis and extends the discussion to deal with correlated variables.

2 Probability Concepts

While it is assumed that the reader has a basic familiarity with statistics, this section is written to present some of the foundational equations of probability theory in which the simulation and interpretation of uncertainty in traffic crash reconstructions is based. The formulas in this section can be found in many texts addressing probability and/or reliability [15–18].

2.1 Multivariate Probability

The usefulness and power of probability and statistics goes well beyond the realm of a single variable. The use of multivariate distributions are useful in modeling and understanding related phenomena.

2.1.1 Joint Correlated Distributions

A joint distribution represents the relationships of the probability of two or more variables. In the case of two random variables, the joint cumulative distribution function (CDF) is defined as

\[
F_{X,Y}(x,y) = P(X \leq x, Y \leq y)
\]

which says the probability of the random variable \(X\) being less than or equal to a particular value of \(x\) and \(Y\) being less than or equal to a particular value of \(y\) follows the joint cumulative distribution function. The cumulative distribution function has a value between 0 and 1 and always has a non-negative
gradient. When dealing with more than two variables, often vector (bold face) notation is used:

\[ F_X(x) = P(X \leq x) \]  

In the same light as the univariate distributions, the joint probability density function (PDF) is defined as

\[ f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} \]  

In the case of a joint normal distribution, the CDF is not defined closed form. However, the multivariate normal probability density function is

\[ f_X(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \bar{\mu})^T \Sigma^{-1} (x - \bar{\mu}) \right] \]  

where \( X \) is a random vector of length \( n \),
\( \Sigma \) is the covariance matrix (\(|\Sigma|\) is the determinant of the covariance matrix),
\( \bar{\mu} \) is the vector of means.

The covariance matrix has the following form

\[ \Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{1,2} \sigma_1 \sigma_2 & \cdots & \rho_{1,n} \sigma_1 \sigma_n \\ \rho_{2,1} \sigma_2 \sigma_1 & \sigma_2^2 & \cdots & \rho_{2,n} \sigma_2 \sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} \sigma_n \sigma_1 & \rho_{n,1} \sigma_n \sigma_2 & \cdots & \sigma_n^2 \end{bmatrix} \]  

where \( \sigma_i \) is the standard deviation of the \( i \)th normal random variable,
\( \sigma_i^2 \) is the variance of the \( i \)th normal random variable,
\( \rho_{i,j} \) is the correlation coefficient between the \( i \)th and \( j \)th variable. (Note: when \( i = j \), \( \rho = 1 \)).

For the case of two random normal (Gaussian) variables, \( X \) and \( Y \), Eq. (4) simplifies to the bivariate normal distribution:

\[ f_{X,Y}(x,y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1 - \rho^2}} \exp \left[ -\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} - \frac{2\rho xy}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2} \right) \right]. \]  

When two variables are statistically independent, then \( \rho = 0 \) and the bivariate distribution is simply the product of two univariate normal distributions. When no correlation is present, as shown in Fig. 1a, the contours are proportional in each direction. If the variates are standardized (i.e. \( z = \frac{x - \mu}{\sigma} \)) then the contours of the distribution surface are concentric circles. When correlation exists, the distribution is oblong as shown in Fig. 1b.

The correlation coefficient is calculated as

\[ \rho = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y} \]  

\[ (7) \]
and has a value from -1 to 1 that indicates the strength of the linear relationship. The covariance\(^1\) in the numerator, \(\text{Cov}(X, Y)\), is defined by the expectation of the product of the deviation in \(X\) and the deviation in \(Y\)

\[
\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - E[X]E[Y] \tag{8}
\]

where the expectation operator is

\[
E[g(x)] = \int_{-\infty}^{+\infty} g(x)f_X(x) \, dx \tag{9}
\]

and is used extensively in applications of probability theory. A correlation coefficient close to 1 or -1 suggests a stronger linear relationship of the data. Similarly, a correlation of 0 suggests no linearity exists in the data.

It is important to note that statistical independence implies no correlation, but the converse is not true. For example, data that fit a circle on the X-Y plane are not correlated, but they are dependent on one another. Also, correlation is strictly a function of the data and does not, by itself, mean a causal relationship exists. Often there is an underlying related causal variable common to both data sets of interest when correlation exists.

---

\(^1\)Another parameter, the coefficient of variation, uses the same abbreviation: \(c.o.v. = \sigma/\mu\).
2.1.2 Conditional and Marginal Distributions

If a set of variables show correlation, then the probability of a value of \( x \) depends on the value of \( y \). This is known as conditional probability and is expressed as

\[
f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}
\]

which says the conditional probability of \( x \) given a value of \( y \) is the ratio of the joint probability to the marginal probability. The denominator in the conditional probability equation is the marginal distribution function and is determined as

\[
f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) \, dx
\]

This marginal distribution is the result of integrating out all the other variables, or flattening all the data onto one axis. A graphical illustration of the concepts of joint, marginal, and conditional PDFs is shown in Fig. 2.

In computing the curves and surfaces in Fig. 2, five pieces of information are given: the two means, two standard deviations, and correlation. These five data completely define a bivariate normal distribution as shown in Fig. 2a. In a computer, the bivariate distribution is typically stored as a 2-D array. The marginal distributions shown in Fig. 2b are symbolically calculated according to Eq. 11 for \( Y \) and

\[
f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) \, dy
\]

for the \( X \) variable. However, when the joint PDF is stored as an array, an approximate joint distribution is determined by summing either the rows or the columns, depending on how the array is set up, and then multiplying by the variable spacing. For example, to sum across all columns and generate a marginal PDF in MATLAB, the code may be \( f_Y = dx \times \text{sum}(f_{XY}, 2) \). For estimating the other marginal PDF, the code would be \( f_X = dy \times \text{sum}(f_{XY}, 1) \).

The conditional PDFs in Figs. 2c & d are estimated by locating the row or column in the joint PDF array that is of interest. Then, simply divide all the values of that row or column by the value of the marginal PDF at the point of interest according to Eq. (10). This division maintains the law of total probability which requires the area under a PDF curve to equal 1. Some example MATLAB code for this operation is as follows:

```matlab
x_interest = 4
I = find(x <= (x_interest + dx/2) & x > (x_interest - dx/2))
fY_givenX = FXY(I,:) / fX(I)
```

It is important to see difference in the conditional distribution in the presence of correlation. However, if no correlation exists, then the conditional and the marginal distributions are exactly the same since the joint distribution is simply the product of all the independent distributions. This can be seen by manipulating...
Figure 2: Example of the concepts of joint (surface), marginal (curves on the edges) and conditional distributions (curves in the middle of the domain). All curves and surfaces are normalized.
Eq. (10) when the conditional and marginal distributions are equal:

\[ f_X(x) \cdot f_Y(y) = f_{X,Y}(x,y) \]  

(13)

2.2 Functions of Random Variables

The goal of an analysis involving random variables is to map random inputs to random outputs through some sort of functional mapping as shown in Fig. 3. In general, the result is a function of several random variables and a closed form solution exists. Consider the result of a function of a random vector:

\[ Z = g(X) \]  

(14)

The distribution of the random vector is a joint distribution \( f_X(x) \) and the goal is to determine the distribution of the results vector, \( f_Z(z) \). If unique inverses exist, then

\[ f_Z(z) = f_X(x)|J| \]  

(15)

where \( |J| \) is the determinant of the Jacobian matrix:

\[
J = \begin{bmatrix}
\frac{\partial x_1}{\partial z_1} & \frac{\partial x_2}{\partial z_1} & \cdots & \frac{\partial x_n}{\partial z_1} \\
\frac{\partial x_1}{\partial z_2} & \frac{\partial x_2}{\partial z_2} & \cdots & \frac{\partial x_n}{\partial z_2} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial x_1}{\partial z_n} & \frac{\partial x_2}{\partial z_n} & \cdots & \frac{\partial x_n}{\partial z_n}
\end{bmatrix}
\]  

(16)

This technique is difficult to implement because the joint distribution is often difficult to write in an analytical form or is simply unknown and must be inferred from the data. As a result, numerical simulation and, when appropriate, the algebra of normal variables is attractive and useful.

2.2.1 Linear Combinations

Of particular importance due to its common application is the fact that a linear combination of normally distributed random variables results in a normal random variable. In general this says that if

\[ Z = \sum_{i=1}^{n} a_iX_i \]  

(17)

and \( X_i \sim N(\mu_i, \sigma_i) \), then \( Z \) is also normally distributed with a mean \( \mu_c \) and standard deviation \( \sigma_c \). Even if the underlying distributions are not known exactly, it can be shown that the resulting mean and standard deviation of the result is known without approximation. In general,

\[ \mu_c = \sum_{i=1}^{n} a_i \mu_{x,i} \]  

(18)
\[ y = f(x) \]

\[ \text{slope} = \text{sensitivity} = \frac{\partial y}{\partial x} \]

Figure 3: The functional mapping of an input to a result and the corresponding uncertainty from a linear approximation as opposed to complete functional mapping.

which is an intuitive result, and

\[ \sigma_z^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{Cov}(X_i, X_j) \tag{19} \]

to determine the variance of the result. In matrix form,

\[ \sigma_z^2 = \mathbf{a}^T \Sigma \mathbf{a} \tag{20} \]

where \( \Sigma \) is the covariance matrix from Eq. (5). This linear combination depends on the correlation between the variables. The case for the linear combination of two random variables (X and Y) is as follows: given

\[ Z = aX + bY \tag{21} \]

\[ \mu_z = a\mu_x + b\mu_y \tag{22} \]

\[ \sigma_z = \sqrt{(a\sigma_x)^2 + (b\sigma_y)^2 + 2ab\rho \sigma_x \sigma_y} \tag{23} \]

### 2.2.2 Taylor Series Approximation

When performing an analysis of an arbitrary function containing several random variables \( Z = g(\mathbf{X}) \), it is possible to obtain approximate solutions to the variance of the result using a Taylor series approximation about the mean values. The Taylor series approximation for a function of two variables is

\[
Z = g(\bar{\mathbf{x}}) + \sum_{i=1}^{n} (X_i - \mu_{x,i}) \frac{\partial g}{\partial X_i} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - \mu_{x,i})(X_j - \mu_{x,j}) \frac{\partial^2 g}{\partial X_i \partial X_j} + H.O.T.
\]
where all derivatives are evaluated at the mean values and \( H.O.T. \) represents all higher order terms. It can be shown that by using the Taylor series with the expectation operator gives an approximation for the variance:

\[
\sigma_z^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial g}{\partial X_i} \bigg|_{\bar{x}} \frac{\partial g}{\partial X_j} \bigg|_{\bar{x}} \text{Cov}(X_i, X_j) \tag{24}
\]

when the random inputs are statistically independent, the covariance matrix becomes diagonal with the individual variances and Eq. (24) becomes the familiar propagation of uncertainty equation:

\[
\sigma_z^2 = \sum_{i=1}^{n} \left( \frac{\partial g}{\partial X_i} \bigg|_{\bar{x}} \right)^2 \sigma_{X_i}^2 \tag{25}
\]

Often the gradients are difficult to obtain in a complex analysis and numerical techniques such as finite differences are used to obtain the gradients.

### 2.2.3 Test for Statistical Independence

Statistical independence can be determined according to the following statistic that follows the student-t distribution [19]:

\[
t = \frac{|\rho| \sqrt{N-2}}{\sqrt{1-\rho^2}} \tag{26}
\]

When \( \rho \) is small, the \( t \) value is smaller and would be rejected by a t-test. As the value of \( \rho^2 \) approaches 1, the value of \( t \) will increase and become significant.

### 2.3 The Monte Carlo Method

The Monte Carlo method is a numerical integration technique for the simulation of probabilistic functions. The technique has a storied history dating back to the Manhattan Project [20]. Its use in traffic crash reconstruction has been accepted and widely publicized [8–14]; however, few have addressed issues related to correlation.

The Monte Carlo Method involves the following steps:

1. Determine probability distributions for the input variables for an analysis

2. Repeatedly sample the input distributions, perform a deterministic analysis, and store the results.

3. Use the stored results to generate an empirical distribution and perform subsequent interpretation.

Often the issue of the number of samples is raised. Since the Monte Carlo method is probabilistic in nature, each attempted run is going to produce different results. Ang and Tang [16] present the following formula to
evaluate the % error in the resulting mean value, \( \bar{z} \):

\[
\%error = 200 \sqrt{\frac{1 - \bar{z}}{n \bar{z}}} \tag{27}
\]

The error decreases as the square root of \( n \) so a reduction of error in half requires four times the number of samples. This can lead to intractable analysis if high precision is needed. As a result, variance reduction techniques such as Latin Hypercube Sampling has become popular [21, 22].

### 3 Modeling and Simulating Correlated Normal Variables

#### 3.1 Theory

Following the instruction found in Refs. [16, 17], the following procedure will generate two correlated random variables following a normal distribution.

1. Generate two independent random values from a standard normal distribution, \( V_1 \sim N(0, 1) \) and \( V_2 \sim N(0, 1) \). While the distributions are the same, the actual arrays of sampled values will be different and independent.

2. Generate a covariance matrix \( \Sigma \) and perform a Cholesky decomposition to get a lower and upper triangular matrix. Cholesky decomposition can be thought of as taking the square root of a matrix.

\[
\Sigma = LL^T \tag{28}
\]

3. Transform the independent standard normal samples to correlated samples with the appropriate variances according to

\[
mU = L^T V \tag{29}
\]

4. Adjust the correlated random variables \( U = [U_1 \; U_2]^T \) to the correct mean value by adding the corresponding mean.

Please note that the above procedure is for normal variables only. It is possible to use different distributions as discussed in [17].

Using this procedure, the probabilistic simulation of energy from crush is presented as an example.

#### 3.2 Example of Determining Energy From Crush

The CRASH3 method has become popular in determining energy from crush [23, 24]. The total crush energy for a perpendicular impact based on an evenly spaced six measurements is determined as:
### Table 1: Damage profiles of the vehicle.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$, inches</td>
<td>0</td>
</tr>
<tr>
<td>$C_2$, inches</td>
<td>5.0</td>
</tr>
<tr>
<td>$C_3$, inches</td>
<td>15.75</td>
</tr>
<tr>
<td>$C_4$, inches</td>
<td>9.0</td>
</tr>
<tr>
<td>$C_5$, inches</td>
<td>1.5</td>
</tr>
<tr>
<td>$C_6$, inches</td>
<td>0</td>
</tr>
<tr>
<td>$L$, inches</td>
<td>61.5</td>
</tr>
</tbody>
</table>

The energy from crush can be calculated using the formula:

$$E_T = \frac{L}{3} \left( \frac{A}{2} (C_1 + 2C_2 + 2C_3 + C_4 + 2C_5 + C_6) \right)$$

$$+ \frac{B}{6} \left[ C_1^2 + 2C_2^2 + 2C_3^2 + 2C_4^2 + C_5^2 + C_1C_2 + C_2C_3 + C_3C_4 + C_4C_5 + C_5C_6 \right]$$

$$+ \frac{5A^2}{2B}$$  \hspace{1cm} (30)

where 
- $L$ is the width of the crush profile,
- $A$ and $B$ are vehicle stiffness coefficients based on crash test data, and
- $C_i$ are evenly spaced crush measurements measured according to the Thumbas and Smith protocol [25].

Many researchers have spent time discussing various aspects of the CRASH3 technique and due to the scatter in vehicle stiffness data a detailed uncertainty analysis is appropriate [26–32].

A crash test at IPTM’s Special Problems conference in April of 2007 had a 1998 Ford Taurus strike a utility pole. The data from that crash were gathered from the scene and vehicle are reported in Table 1.

In addition to knowing the damage profile, the stiffness coefficients must be found. These data were obtained from a searchable database, StiffCALCS, from Expert AutoStats. These data are shown in Table 2. This database enables the investigator to search for applicable tests for a vehicle make and model, exporting all of the data statistically. Since no other information is available, it is assumed that the stiffness values follow a normal distribution.

Using the average stiffness values in Table 2 and the crush measurements of Table 1, the total energy from crush according to Eq. (30) is 42,350 ft-lb. This value is based on empirical results from crash tests and is known to be highly variable. Also, the crush stiffness values are correlated. A plot of the stiffness data is shown in Fig. 4. For this example, only the stiffness values are considered to be random, but a detailed analysis should include the uncertainty in the other input variables as well.
<table>
<thead>
<tr>
<th>Stiffness</th>
<th>A, lb/in</th>
<th>B, lb/in^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>418.4</td>
<td>161.9</td>
</tr>
<tr>
<td>Maximum</td>
<td>486.7</td>
<td>228.6</td>
</tr>
<tr>
<td>Minimum</td>
<td>357</td>
<td>109.6</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>41.9</td>
<td>36.5</td>
</tr>
</tbody>
</table>

Table 2: Frontal stiffness values for Ford Taurus / Mercury Sable vehicles of this body style (1996 - 1999, eleven tests)

B = -178.9851 + 0.81485(A) \quad R^2 = 0.87901

Figure 4: Scatter plot and linear fit of the A and B crush stiffness coefficients showing a positive correlation.
A least square fit of a line, $B = mA + b$ is also shown in Fig. 4. The coefficient of determination is determined as

$$R^2 = 1 - \frac{Var(\varepsilon)}{Var(B)}$$

(31)

where $\varepsilon = B - \hat{B}$ which is the difference between the data and the fitted line. In terms of variance, the correlation coefficient is

$$\rho^2 = 1 - \frac{n - 2 Var(\varepsilon)}{n - 1 Var(B)}$$

(32)

which shows the coefficient of determination is a close predictor for the correlation coefficient and is equal to $R^2$ as $n$ gets large. These parameters are often confused since for a linear fit, the square of the correlation has the nearly same value as the coefficient of determination. However, the interpretation is different. For example, the correlation coefficient is a measure of linearity, but the coefficient of determination is used to assess the quality of the least squares fit— even if the fitted function is not a line.

It should be pointed out that, according to Eq. (26), the stiffness data are significantly correlated with a correlation coefficient of $\rho = 0.9376$ and the t statistic of 8.086. This suggests that the probability of the shown linearity in the data due to random chance is extremely small.

Simulating the stiffness values requires the covariance matrix $\Sigma$ which is determined using built in functions in Matlab. The results of $\text{cov}(A, B)$ are

$$\text{cov}(A, B) = \begin{bmatrix} 1759.2 & 1433.5 \\ 1433.5 & 13289 \end{bmatrix}$$

and its corresponding Cholesky decomposition is

$$\text{chol}(\text{cov}(A, B)) = \begin{bmatrix} 41.9431 & 34.1773 \\ 0 & 12.6802 \end{bmatrix}$$

The Cholesky decomposition of the covariance matrix can be thought of as determining the “standard deviation matrix” in a loose sense. With these arrays defined, the Monte Carlo simulation can commence. The results of both independent sampling and correlated sampling are shown in Fig. 5.

The simulated data can be displayed in Fig. 6 using a 3-D histogram in addition to the scatter plots of Fig. 5. These figures are proportional to the bivariate normal distributions shown in Fig. 2.

Based on the correlated sampling, the energy can be computed according to Eq. (30) and the results are represented as a histogram shown in Fig. 7. While the histogram is useful in representing the results, the probabilities and ranges are more obvious from the CDF, as shown in Fig. 8. It can be seen, both from this example and Eq. (24), that positive correlation tends to increase the variance of a result. The reason for this is because the correlated samples influence each other in that when a high value of the first variable is chosen it is more likely that a high value from the second distribution is sampled. For this case, the consequence of two high stiffness values is a high energy result. Likewise, the combination of low values decreases the energy. The overall effect is to widen the probability distribution and make the extreme values of energy
Figure 5: The sampling results superimposed on the actual data for the A and B stiffness values.

Figure 6: Three dimensional histograms for the simulated stiffness values. These plots duplicate the data shown in Fig. 5.
The effect of negative correlation is also shown in this example in Fig. 8. With the input variables negatively correlated, a high sample from the first random variable with more likely produce a low value in the second. When discussing energy from crush, these samples tend to offset each other and the resulting distribution is narrower. It is important to realize that the functional relationship will determine the effect of the correlation in terms of adjusting the variance of the final distribution.

4 Predicting Correlation

In some cases, such as a linear combination, there are analytical methods for predicting correlation. A situation where this is applicable is in measurement theory. When measuring something with an instrument, there are two types of uncertainty: aleatory and epistemic. Definitions are as follows [16, 33]:

**Aleatory** uncertainty comes from the inherent variability of a system. It can be deduced though repeated measurements and cannot be reduced. Its synonyms include random and precision uncertainty.

**Epistemic** uncertainty comes from a lack of information. This sometimes is known as modeling uncertainty, systematic uncertainty, or bias. In measurement systems it can be reduced though thoughtful calibration. Often times it is an estimated quantity based on past experience or engineering judgment.

Consider measuring two different objects with the same instrument. For example, weigh two vehicles that have been in an accident on the same scale. The instrument itself has a tolerance and a certain amount of random uncertainty experienced between different measurements. The random uncertainty associated with
Figure 8: Cumulative distributions showing the effect of correlation (both positive and negative) compared to the effect of no correlation. The horizontal bars indicate the differences in the 95% most likely probability range. The mean response is the same for all three cases.
two different measurements is denoted as $P_1$ and $P_2$. These values may have the same value but are considered independent.

Also, the instrument has some systematic uncertainty associated with it due to calibration, field use, temperature, etc. We consider this value to be the same between measurements and is denoted as $B$. Therefore, the two measured values, which are considered random variables, are modeled as a linear combination:

\begin{align*}
R_1 &= \mu_1 + P_1 + B \quad (33) \\
R_2 &= \mu_2 + P_2 + B \quad (34)
\end{align*}

Since both measurement results $R_1$ and $R_2$ have a common random term, they are necessarily correlated. Presumably the variances of $P_1$, $P_2$, and $B$ are known and the correlation between $R_1$ and $R_2$ is desired. To determine the correlation, the mean values are irrelevant and a modification of Eq. 23 is used

\begin{align*}
Var(R_1) &= Var(P_1) + Var(B) \quad (35) \\
Var(R_2) &= Var(P_2) + Var(B) \quad (36)
\end{align*}

because $P$ and $B$ are independent. Now consider the difference of the two results

\[ R_{\text{diff}} = R_1 - R_2 = \frac{\mu_1 - \mu_2 + P_1 - P_2}{\text{constant}} \quad (37) \]

Since $R_1$ and $R_2$ are correlated, the variance of $R_{\text{diff}}$ is

\[ Var(R_{\text{diff}}) = Var(R_1) + Var(R_2) - 2Cov(R_1, R_2) \quad (38) \]

rearrange and make the appropriate substitutions gives the covariance as:

\[ Cov(R_1, R_2) = \frac{Var(R_1) + Var(R_2) - Var(R_{\text{diff}})}{2} \quad (39) \]

but since $P_1$ and $P_2$ are independent,

\[ Var(R_{\text{diff}}) = Var(P_1) + Var(P_2) \quad (40) \]

after making the appropriate substitutions

\[ Cov(R_1, R_2) = \frac{Var(P_1) + Var(B) + Var(P_2) + Var(B) - Var(P_1) - Var(P_2)}{2} \quad (41) \]

which simplifies to

\[ Cov(R_1, R_2) = Var(B) \quad (42) \]
from which the correlation coefficient can be determined from Eq. (7). This final calculation, of course, needs the standard deviations of the two measurement results which come from Eqs. (35) and (36). The random uncertainty $P_i$ comes from repeated measurement. With low resolution instrumentation, sometimes the repeated measurement reveal no random uncertainty, which would suggest the two measurements to be perfectly correlated.

This perfect correlation assumption may not be applicable in that the epistemic uncertainty could be broken into a correlated and an uncorrelated part. For example, consider weighing two different cars in an attempt to determine the vehicle weights during a crash. If a large platform scale is used, the displayed weight will not likely reveal any fluctuations between measurements thus suggesting no random uncertainty. However, there will be some imperfect correlation between the vehicles as the weight at the time of the crash includes the current vehicle weight, the weight of the driver and passengers, the weight of cargo that has been removed and the weight of fluids and car parts discarded since the time of the crash. The weight of the $i$th vehicle could be written mathematically as a sum:

$$w_i = \mu_i + w_{i,people} + w_{i,cargo} + w_{i,debris} + w_{i,fluids}$$

where $\mu_i$ is the measured weight of the vehicles. Often even the measured weight must be determined though a database or other indirect means. When the different weight components are estimated, rather than measured, then the corresponding uncertainty is higher.

When the same scale is used and the same expert is estimating the uncertainties of the missing weight components, the correlation between the two vehicles in a crash is inevitable. However, it is apparent that variation between the weight estimates of the cars exist as well. Therefore, in the absence of better information, the author recommends using a correlation coefficient between the estimated weights of two vehicles involved in a crash to be 0.8.

5 Interpreting Correlated Results from a Monte Carlo Simulation

Often a Monte Carlo simulation produces more than one output parameter which are likely to be dependent on one another. Therefore, the purest numerical result is the estimated joint distribution. However, multi-dimensional joint distributions are difficult to display and interpret, so marginal distributions are commonly used to display the results.

The sole use of marginal distributions can lead to misleading results, especially reporting opinions on likely scenarios for a crash. Instead, results should be reported as conditional distributions so as to demonstrate that the results are related.

For example, consider a planar conservation of linear momentum analysis of two point masses (ignore rotation and external impulses). The well known procedure takes known exit velocities, entrance angles, departure angles, and weights to compute the speeds at impact. Often there are four results reported in a momentum analysis: $v_1$, $v_2$, $\Delta v_1$, and $\Delta v_2$. For this example, consider only the entrance speeds as calculated
Table 3: Monte Carlo Simulation Inputs for an oblique impact.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Symbol</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Distribution</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight 1 (lb)</td>
<td>$w_1$</td>
<td>3215</td>
<td>58</td>
<td>Bivariate Normal</td>
<td>$\rho_{w_1,w_2} = 0.67$</td>
</tr>
<tr>
<td>Weight 2 (lb)</td>
<td>$w_2$</td>
<td>4320</td>
<td>64</td>
<td>Bivariate Normal</td>
<td>$\rho_{w_2,w_1} = 0.67$</td>
</tr>
<tr>
<td>Approach Angle 2 (deg)</td>
<td>$\psi$</td>
<td>110</td>
<td>3</td>
<td>Normal</td>
<td>0</td>
</tr>
<tr>
<td>Exit Angle 1 (deg)</td>
<td>$\theta$</td>
<td>30</td>
<td>2</td>
<td>Normal</td>
<td>0</td>
</tr>
<tr>
<td>Exit Angle 2 (deg)</td>
<td>$\phi$</td>
<td>60</td>
<td>2</td>
<td>Normal</td>
<td>0</td>
</tr>
<tr>
<td>Exit Speed 1 (mph)</td>
<td>$v_3$</td>
<td>23.2</td>
<td>2</td>
<td>Normal</td>
<td>$\rho_{w_1,w_2} = 0.87$</td>
</tr>
<tr>
<td>Exit Speed 2 (mph)</td>
<td>$v_4$</td>
<td>35.4</td>
<td>3</td>
<td>Normal</td>
<td>$\rho_{w_2,w_1} = 0.87$</td>
</tr>
</tbody>
</table>

with the following formulas [24, 34]:

$$v_2 = \frac{w_1 v_3 \sin \theta}{w_2 \sin \psi} + \frac{v_4 \sin \phi}{\sin \psi}$$  \hspace{1cm} (43)

$$v_1 = v_3 \cos \theta + \frac{w_2 v_4 \cos \phi}{w_1} - \frac{w_2 v_2 \cos \psi}{w_1}$$  \hspace{1cm} (44)

Notice, for any impact angle $\psi$ other than 90 degrees, the result of $v_1$ depends on the value of $v_2$. Furthermore, the weights, angles, and exit velocities may be correlated in some manner. For example, if both vehicles slide or spin to a stop on the same surface, then the exit velocities will be correlated. A momentum analysis of this nature lends itself well to Monte Carlo Simulation and an example of repeatedly simulating a momentum problem is shown in Fig. 9. If all input variables were to be treated as random, then the correlated random samples would be generated by using an array of independent standard normal samples of size 7 by $n$. The 7 by 7 covariance matrix will have the information for the correlations. If two variables are independent, then the off-diagonal term of the covariance matrix for the resulting correlation will have a zero. If all inputs are independent, then the covariance matrix will be diagonal.

If only two results for the momentum analysis are considered, the results of the Monte Carlo simulation can be displayed as a 3-D histogram and interpreted using the concepts and procedures of Section 2. The following numerical example shows this procedure and interpretation of results.

Consider an oblique impact and the momentum solution the seven input variables and their corresponding input distributions are given in Table 3. The solution to the momentum problem in terms of the mean values gives $v_1 = 80.5$ mph and $v_2 = 41.8$ mph. However, the results of the Monte Carlo analysis, $v_1$ and $v_2$, comprise a joint distribution which is estimated as a bivariate PDF shown in Fig. 10. If the list of stored results are treated independently, the resulting estimated distributions would be the marginal distributions shown in Fig. 11.

When reporting final results from an analysis, a range of the results are often presented. However in the case of correlated outputs, the results of one variable depend on the results of the other and the final results must be interpreted from conditional probabilities. This means multiple ranges should be reported for the results. For example, if we take the 90% range as our precision interval from the marginal distribution,
Figure 9: The vector diagram solution to the conservation of momentum analysis showing the iterative nature of the sampling scheme. Notice how all the results are related which means the output of the probabilistic analysis will be a correlated joint distribution.
Figure 10: The estimated bivariate distribution of the Monte Carlo Simulation results.

Estimated Joint distribution (μₓ = 80.4617, μᵧ = 41.8346, αₓ = 5.939, αᵧ = 3.0199, and ρ = 0.86769)
Figure 11: Marginal distributions showing the distribution shape of the results one at a time. These marginal distributions show the probability for all possibilities other variables. The distributions have been normalize to have a total area of 1.
then the range of \( v_2 \) is from 36.97 mph to 46.62 mph. As shown in Fig. 12, the conditional distribution associated with each bound is significantly different than the overall distribution. This can lead to significant and, perhaps, misleading interpretation of the Monte Carlo results.

The 90% range of \( v_1 \) given that \( v_2 \) is 36.97 mph is between 63.42 and 73.07 mph. Also, the 90% range of \( v_1 \) given that \( v_2 \) is 46.62 is between 85.75 and 95.90 mph. In a similar fashion, the conditional probabilities of \( v_2 \) given extreme values in \( v_1 \) are shown in Fig. 12 as

\[
\begin{align*}
v_{1,\text{min}} &= 70.9 \text{ mph} \\
v_{2,\text{min}} &= 37.5 \text{ mph} \\
v_{2,\text{max}} &= 41.3 \text{ mph}
\end{align*}
\]

and

\[
\begin{align*}
v_{1,\text{max}} &= 90.3 \text{ mph} \\
v_{2,\text{min}} &= 42.8 \text{ mph} \\
v_{2,\text{max}} &= 47.2 \text{ mph}
\end{align*}
\]

The results depicted in Fig. 12 show, for this example, that including the conditional probabilities reveal different ranges of results. The positive correlation shows that it is more likely that both vehicles were traveling on the upper level of their speed ranges or on the lower level of their speed ranges. The combination of high/low or low/high is much less likely. Using conditional probabilities also reduces the ranges calculated from the distribution. This overcomes some of the criticism that Monte Carlo results produce an unusable range of results. Conditional probabilities also enable the analyst to evaluate the opposition’s results with a tighter tolerance to detect any inconsistencies.

6 Conclusions

The goal of this paper was to show the effect of including correlation on uncertainty analysis using traffic crash reconstruction examples. The mathematical basis for joint, marginal, and conditional probability distributions for continuous variables was briefly reviewed. Uncertainty analysis techniques, including analytical solutions, Taylor series approximations, and the Monte Carlo method, were discussed in the context of correlated variables. An example of generating and simulating correlated crush stiffness values for a Monte Carlo simulation of crush energy determination using CRASH3 showed the uncertain of the energy associated with residual crush increased when a positive correlation exists between A and B stiffness values. An analytical demonstration of determining the correlation and covariance of the measured values of different things using the same instrument was shown with application to weighing vehicles. It was also revealed that multiple results of an analysis can be correlated and interpreting those results requires using the concepts of multivariate probability. A solution to a crash reconstruction using the conservation of linear momentum was presented to demonstrate how the final speed ranges can be reduced by examining the conditional probability. The effect of correlation can be significant and uncertainty analysis in traffic crash reconstruction and the analyst should address issues of correlation when developing inputs for an uncertainty analysis as well as be able to
Figure 12: The conditional distributions stemming from the joint Monte Carlo simulation results.
interpret correlated results.

References


