

Monte Carlo Techniques for Correlated Variables in Crash Reconstruction

Jeremy S. Daily
The University of Tulsa

Copyright © 2009 SAE International

ABSTRACT

The results of a traffic crash reconstruction often include vehicle speeds to address causation and changes in velocity to indicate crash severity. Since these results are related, they should be modeled in a probabilistic context as a joint distribution. Current techniques in the traffic crash reconstruction literature assume the input parameters and results of an analysis are independent, which may or may not be appropriate. Therefore, a discussion of uncertainty propagation techniques with correlation and Monte Carlo simulation of correlated variables is presented in this paper. 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 correlated. Results show the difference between accounting for correlation and assuming independence may be significant. However, the examples provided are aimed at introducing the concept of correlation in Monte Carlo simulation and determining the practical significance of correlation have yet to be determined. 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.

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 a precise result. As such, the results of every crash analysis must include some measure indicating the reliability (uncertainty) of the result. To this end, many authors have presented techniques for dealing with uncertainty in crash reconstruction. [1–7].

The range of probable analysis results 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 address 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. 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 use simple equations that are tractable by hand, computer based reconstruction programs can also benefit from the concept dealing with correlation in an uncertainty analysis. Finally, it should be pointed out that the practical significance of the effects of correlation in crash reconstruction has yet to be determined since the examples provided herein are just aimed at introducing the concepts of correlation.

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].

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.

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

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) \quad (1)$$

which says the probability of the random variable X being less than or equal to a particular value of x and the random variable 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:

$$F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) \quad (2)$$

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

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} \quad (3)$$

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

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \bar{\mu})^T \Sigma^{-1} (\mathbf{x} - \bar{\mu}) \right] \quad (4)$$

where $\mathbf{X}$ is a random vector of length n , Σ is the covariance matrix ($|\Sigma|$ is the determinant of the covariance matrix),

and $\bar{\mu}$ is the vector of means. The covariance matrix has the following form

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{1,2} \sigma_1 \sigma_2 & \cdots & \rho_{1,n} \sigma_1 \sigma_n \\ \rho_{2,1} \sigma_2 \sigma_1 & \sigma_2^2 & \cdots & \rho_{2,n} \sigma_2 \sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} \sigma_n \sigma_1 & \rho_{n,2} \sigma_n \sigma_2 & \cdots & \sigma_n^2 \end{bmatrix} \quad (5)$$

where σ_i is the standard deviation of the i th normal random variable, σ_i^2 is the variance of the i th normal random variable, and $\rho_{i,j}$ 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:

$$f_{X,Y}(x,y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} \exp \left[\frac{-1}{2(1-\rho^2)} \left(\frac{x^2}{\sigma_x^2} - \frac{2\rho xy}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2} \right) \right] \quad (6)$$

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 the contours of the probability density function are proportional in each direction, as shown in Fig. 1a. 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 contours are oblong with a tilted major axis, as shown in Fig. 1b.

The correlation coefficient is calculated as

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y} \quad (7)$$

and has a value from -1 to 1 that indicates the strength of the linear relationship. The covariance¹ 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

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

where the expectation operator is

$$E[g(\mathbf{x})] = \int_{-\infty}^{+\infty} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \quad (9)$$

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

¹ Another parameter, the coefficient of variation, uses the same abbreviation: *c.o.v.* = σ/μ .

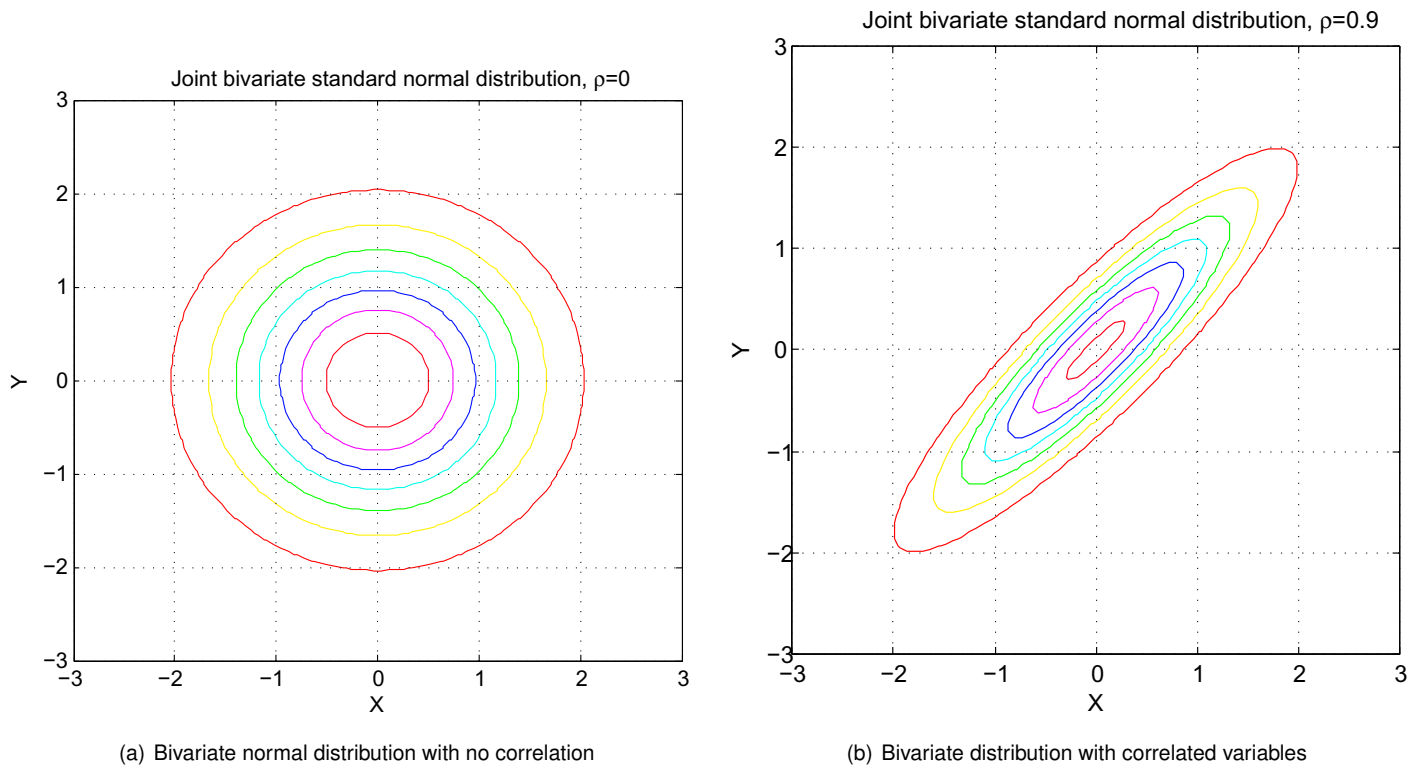


Figure 1: The standard normal joint distribution contours.

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.

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

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad (10)$$

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

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dx \quad (11)$$

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

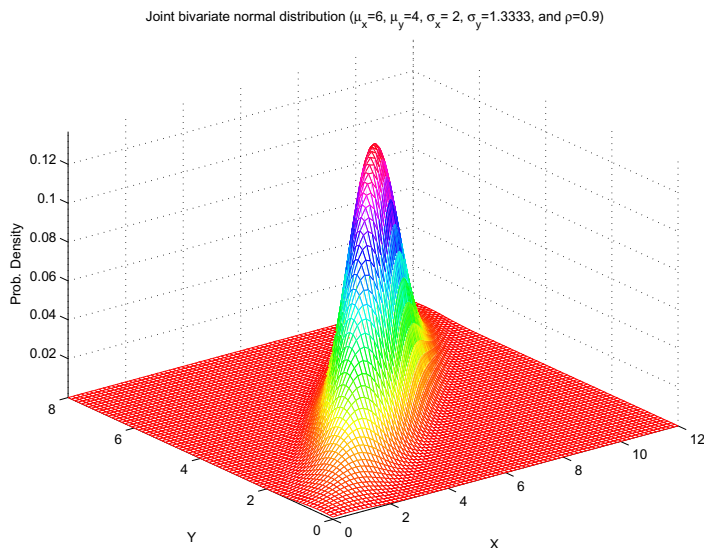
$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy \quad (12)$$

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 `fY=dx*sum(fXY,2)`. For estimating the other marginal PDF, the code would be `fX=dy*sum(fXY,1)`.

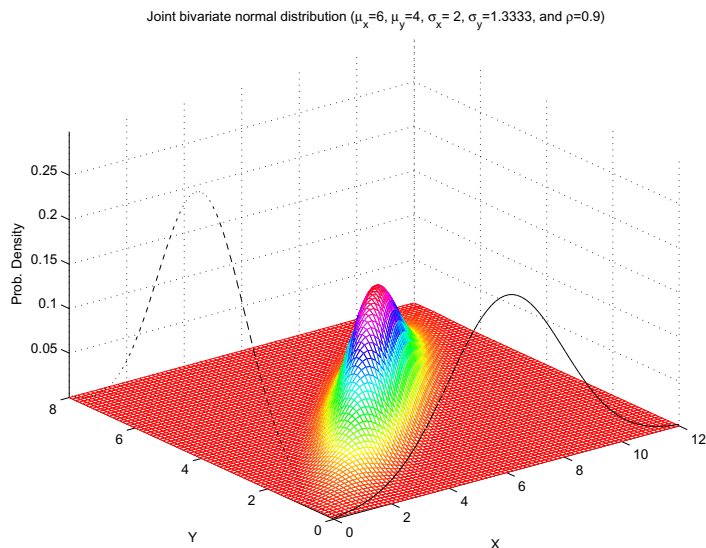
The conditional PDFs in Figs. 2c & d are estimated by locating the row or column in the joint PDF array that is of interest. Then, simply divide all the values of that row or column by the value of the marginal PDF at the point of interest according to Eq. (10). This division maintains the law of total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

There is a difference between the conditional distributions 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 Eq. (10) when the conditional and marginal distributions are equal:

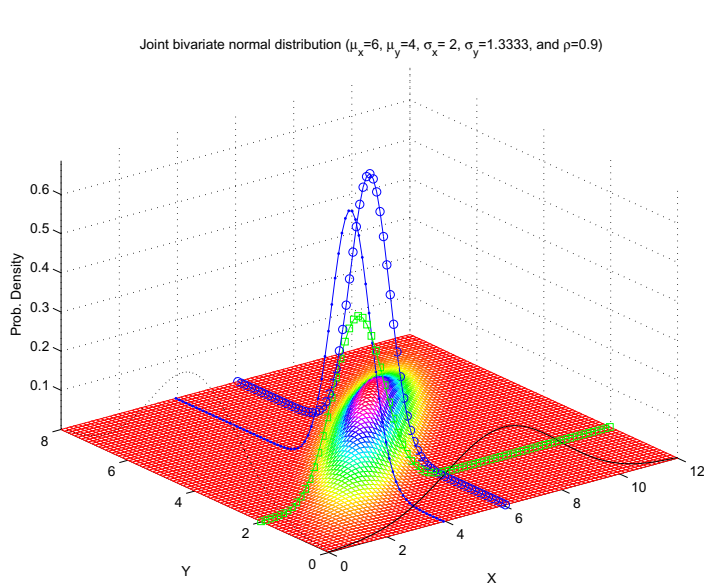
$$f_X(x) \cdot f_Y(y) = f_{X,Y}(x,y)$$



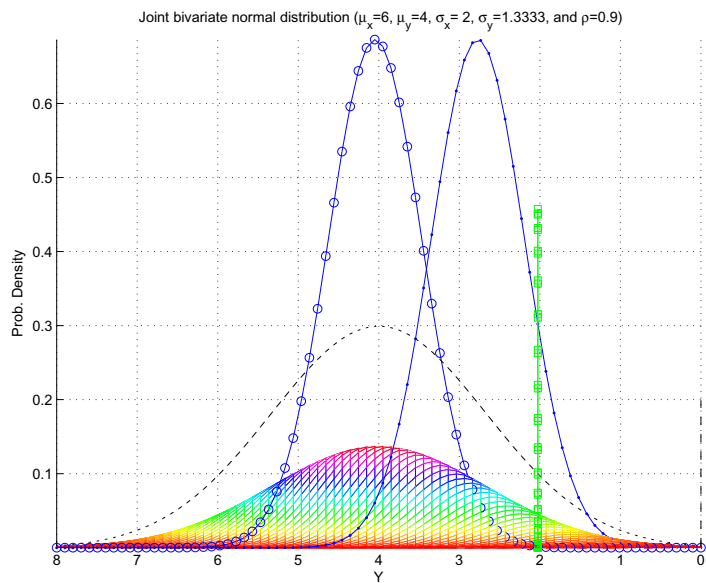
(a) The surface of a correlated joint bivariate distribution



(b) The marginal distributions are the results of an integration



(c) Conditional probability density functions



(d) The conditional probabilities, $f_{Y|X}(y, x)$, produce different results depending on the choice of x .

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.

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 may exist. Consider the result of a function of a random vector:

$$\mathbf{Z} = g(\mathbf{X}) \quad (14)$$

The distribution of the random vector is a joint distribution $f_{\mathbf{X}}(\mathbf{x})$ and the goal is to determine the distribution of the results vector, $f_{\mathbf{Z}}(\mathbf{z})$. If unique inverses exist, then

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(\mathbf{x})|J| \quad (15)$$

where $|J|$ is the determinant of the Jacobian matrix:

$$J = \begin{bmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_2}{\partial z_1} & \cdots & \frac{\partial x_n}{\partial z_1} \\ \frac{\partial x_1}{\partial z_2} & \frac{\partial x_2}{\partial z_2} & \cdots & \frac{\partial x_n}{\partial z_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_1}{\partial z_n} & \frac{\partial x_2}{\partial z_n} & \cdots & \frac{\partial x_n}{\partial z_n} \end{bmatrix} \quad (16)$$

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 is attractive and useful for determining the distribution of functions of random variables.

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

$$Z = \sum_{i=1}^n a_i X_i, \quad (17)$$

a_i are constant coefficients, and $X_i \sim N(\mu_i, \sigma_i)$, then Z is also normally distributed with a mean μ_z and standard deviation σ_z . 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,

$$\mu_z = \sum_{i=1}^n a_i \mu_{x,i} \quad (18)$$

which is an intuitive result, and

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

which is used to determine the variance of the result. In matrix form,

$$\sigma_z^2 = \mathbf{a}^T \Sigma \mathbf{a} \quad (20)$$

where Σ 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:

$$Z = a_1 X + a_2 Y \quad (21)$$

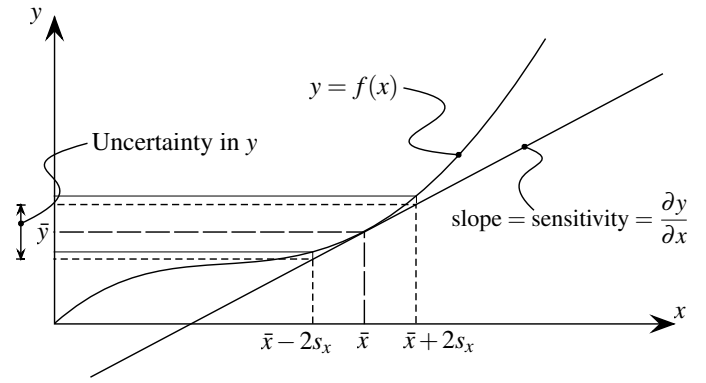


Figure 3: The functional mapping of an input to a result and the corresponding uncertainty from a linear approximation as opposed to complete functional mapping.

$$\mu_z = a_1 \mu_x + a_2 \mu_y \quad (22)$$

$$\sigma_z = \sqrt{(a_1 \sigma_x)^2 + (a_2 \sigma_y)^2 + 2a_1 a_2 \rho \sigma_x \sigma_y} \quad (23)$$

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

$$Z = g(\bar{\mu}_x) + \sum_{i=1}^n (X_i - \mu_{x,i}) \frac{\partial g}{\partial X_i} + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (X_i - \mu_{x,i})(X_j - \mu_{x,j}) \frac{\partial^2 g}{\partial X_i \partial X_j} + H.O.T. \quad (24)$$

where all derivatives are evaluated at the mean values and $H.O.T.$ represents all higher order terms. It can be shown that using the Taylor series with the expectation operator gives an approximation for the variance:

$$\sigma_z^2 = \sum_{i=1}^n \sum_{j=1}^n \left. \frac{\partial g}{\partial X_i} \right|_{\bar{\mu}_x} \left. \frac{\partial g}{\partial X_j} \right|_{\bar{\mu}_x} \text{Cov}(X_i, X_j) \quad (25)$$

when the random inputs are statistically independent, the covariance matrix becomes diagonal with the individual variances and Eq. (25) becomes the familiar propagation of uncertainty equation:

$$\sigma_z^2 = \sum_{i=1}^n \left(\left. \frac{\partial g}{\partial X_i} \right|_{\bar{\mu}_x} \right)^2 \sigma_{X_i}^2 \quad (26)$$

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

Test for Statistical Independence Statistical independence (i.e. insignificant correlation) can be determined according to the following statistic that follows the