Proceedings of the 1993 Winter Simulation Conference
G.W. Evans, M. Mollaghasemi, E.C. Russell, W.E. Biles (eds.)

A COMPARISON OF ALTERNATIVE INPUT MODELS FOR SYNTHETIC OPTIMIZATION PROBLEMS

Charles H. Reilly

Department of Industrial and Systems Engineering
The Ohio State University
1971 Neil Avenue
Columbus, Ohio, 43210, U.S.A.

ABSTRACT

We analyze two strategies for randomly generating optimization test problems with two types of coefficients. One strategy is to generate test problems with independent coefficients; the other strategy is to generate test problems with induced correlation between the coefficient types. We discuss the likely effect of test problem size, i.e., the number of decision variables, on the sample correlations among the test problem coefficients generated with each strategy. We also propose some guidelines for experimenters based on our analysis.

1 INTRODUCTION

Typically, synthetic test problems are randomly generated when an evaluation of solution methods for an optimization problem is conducted because the pool of real-world instances with known optimal solutions is too small to provide enough distinct test cases. Unfortunately, the generation of synthetic problems is rarely viewed as the multivariate sampling problem it truly is.

In order to generate test problems, certain assumptions about the coefficients must be made. For example, a distribution must be specified for each coefficient type, and relationships between the coefficient types and/or the constants in the test problems may also be specified. There are exceptions, but it is common to assume that some or all of the coefficient types are mutually independent and to assume that each coefficient type is uniformly distributed.

The primary motivation for this paper is the random generation of synthetic test problems with two types of coefficients. However, the work that is presented here applies to the more general problem of generating samples of a bivariate discrete random variable with specified marginal mass functions.

In the next section, we review some background material. In §3, we suggest an estimator for the expected correlation between two random variables that is the basis for our later comparison of alternative input models. In §4, we state the assumptions that we make for our analysis and discuss some of the consequences of our assumptions. We consider two strategies for generating test problem coefficients: generating pairs of coefficients under the assumption of independence (§5) and generating coefficient pairs with induced correlation (§6). We address the effect of test problem size on the range of sample correlations between the two types of coefficients and suggest guidelines for test problem dimensions and the number of test problems. We conclude with a brief discussion in §7.

2 BACKGROUND

In this section, we review the concept of a parametric envelope for a bivariate discrete random variable (X, Y), as well as conventional mixtures of bivariate pmfs that are often used to simulate values of (X, Y) when $\rho = \mathrm{Corr}(X, Y)$ is specified. We also discuss some relevant computational experiments on synthetic optimization problems.

2.1 Parametric Envelope for (X,Y)

Let X be a discrete random variable distributed over n_1 values in $S_X = \{x_1, x_2, \ldots, x_{n_1}\}$, where $x_1 < x_2 < \cdots < x_{n_1}$, according to the pmf $f_1(x)$ and Y be a discrete random variable distributed over n_2 values in $S_Y = \{y_1, y_2, \ldots, y_{n_2}\}$, where $y_1 < y_2 < \cdots < y_{n_2}$, according to the pmf $f_2(y)$. A curve that plots θ as a function of ρ, where θ is the largest possible value of the smallest joint probability over the bivariate support $S_X \times S_Y$, can be constructed following the solution of a parametric linear program (Peterson, 1990; Peterson and Reilly, 1991). Peterson and Reilly (1991) and Reilly (1991) suggest that this parametric curve defines a parametric envelope of all feasible combinations of ρ and θ. At least one pmf is associated with each point in the parametric envelope.

We adopt the parametric envelope concept suggested in Peterson and Reilly (1991) and Reilly (1991) and relate it to the random generation of optimization test problems.

2.2 Conventional Mixtures

A common way to characterize the pmf for (X, Y) when $\rho = \rho_0$ is to mix values of (X, Y) generated under the assumption that X and Y are independent and values of (X, Y) generated under the assumption that X and Y have extreme correlation. (For example, see Schmeiser and Lal (1982).) Let $g_{min}(x, y)$ be the minimum-correlation pmf for (X, Y) and $g_{max}(x, y)$ be the maximum-correlation pmf for (X, Y). Conventional mixtures have the form:

$$g_c(x,y) = \begin{cases} \lambda_0^{c+} f_1(x) f_2(y) + \\ \quad \lambda_{max}^{c+} g_{max}(x,y) & \text{if } \rho_0 \geq 0; \\ \\ \lambda_0^{c-} f_1(x) f_2(y) + \\ \quad \lambda_{min}^{c-} g_{min}(x,y) & \text{if } \rho_0 < 0; \end{cases} \tag{1}$$

where $\lambda_{max}^{c+} = \rho_0/\rho_{max}$, $\lambda_0^{c+} = 1 - \lambda_{max}^{c+}$, $\lambda_{min}^{c-} = \rho_0/\rho_{min}$, $\lambda_0^{c-} = 1 - \lambda_{min}^{c-}$, and ρ_{max} and ρ_{min} are the maximum and minimum values possible for ρ, respectively.

Conventional mixtures (1) can be used to generate values of (X, Y) for any feasible ρ_0. However, for each feasible ρ_0, there is a unique conventional mixture. If we are interested in generating values of (X, Y) with X and Y dependent and uncorrelated, a conventional mixture will not suffice.

2.3 Related Computational Experiments

There have been many, many computational evaluations of solution methods conducted. We summarize some of the studies pertinent to our analysis.

Loulou and Michaelides (1979) use two different distributions, the uniform and the 2-Erlang, for the constraint coefficients in multidimensional knapsack problems. They conclude that the statistical properties of test problems can affect the performance of solution methods. Similar conclusions about the effect of correlation have been drawn in other studies (Martello and Toth, 1979; Balas and Martin, 1980; Balas and Zemel, 1980; Potts and Van Wassenhove, 1988; John, 1989; Moore, 1989; Moore, 1990)

Martello and Toth (1979) compare the performance of algorithms for the 0-1 knapsack problem on test problems in which the objective and constraint coefficients are uncorrelated (independent), "weakly correlated", and "strongly correlated" ($\rho = 1$). They vary the correlation between the coefficient types by changing the distribution of the objective function coefficients.

Balas and Zemel (1980) report on an evaluation of solution methods for the 0-1 knapsack problem using the same types of test problems as Martello and Toth. Potts and Van Wassenhove (1988) and John (1989) use a very similar approach to generate synthetic scheduling problems.

Balas and Martin (1980) include capital budgeting test problems in which the constraint coefficients for each variable are related to the objective function coefficient for that variable.

Moore (1989) studies 0-1 knapsack problems that are generated using conventional mixtures (1) in order to assess the effect of the expected correlation between the objective and constraint coefficients on the performance of a simple implicit enumeration routine. Moore (1990) uses conventional mixtures (1) to generate weighted set covering problems in which correlation is induced between the objective function coefficients and the sum of the binary constraint coefficients.

Hoffman and Jackson (1982), Greenberg (1990), and Jackson, Boggs, Nash, and Powell (1991) are examples of papers that provide some guidance for designing and reporting computational experiments.

3 A CORRELATION ESTIMATOR

Let $h(x, y)$ be any valid pmf for (X, Y) and consider the random variable XY. By definition,

$$\mu_{XY|h} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} x_i y_j h(x_i, y_j)$$

and

$$\sigma_{XY|h}^2 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (x_i y_j)^2 h(x_i, y_j) - \mu_{XY|h}^2.$$

Suppose that N observations of (X, Y) are generated using $h(x, y)$, and let

$$R_{N|h} = \frac{\sum_{k=1}^{N} x_k y_k / N - \mu_X \mu_Y}{\sigma_X \sigma_Y}.$$

$R_{N|h}$ is an unbiased estimator of $\rho_h = \text{Corr}(X, Y|h)$. Furthermore, $R_{N|h}$ is asymptotically normally distributed with mean

$$\rho_h = \frac{\mu_{XY|h} - \mu_X \mu_Y}{\sigma_X \sigma_Y}$$

and variance

$$\sigma_{R_{N|h}}^2 = \frac{\sigma_{XY|h}^2}{N \sigma_X^2 \sigma_Y^2}.$$

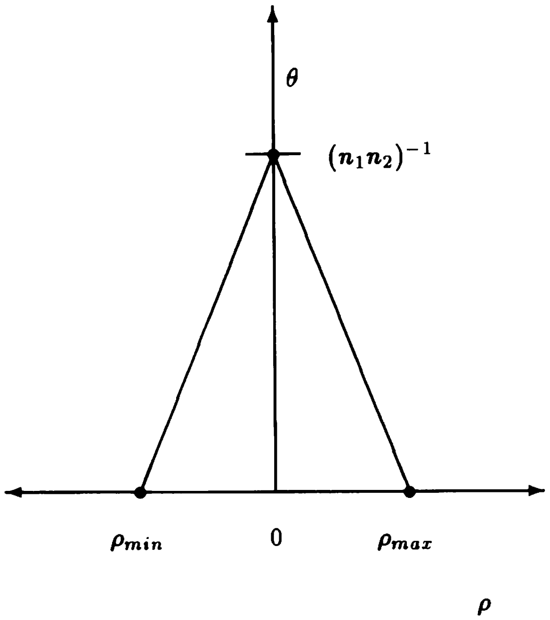

Figure 1: Parametric envelope for (X, Y)

In §5 and §6, we use the estimator $R_{N|h}$ to analyze the generation of values of (X, Y) with independent sampling and sampling with induced correlation, respectively.

4 ASSUMPTIONS

Throughout the remainder of this paper, we assume that X and Y are discrete random variables such that $X \sim U\{1, 2, \ldots, n_1\}$ and $Y \sim U\{1, 2, \ldots, n_2\}$. It follows that $\mu_X = (n_1 + 1)/2$, $\sigma_X^2 = (n_1^2 - 1)/12$, $\mu_Y = (n_2 + 1)/2$, and $\sigma_Y^2 = (n_2^2 - 1)/12$.

Peterson and Reilly (1991) show that when X and Y are uniformly distributed the parametric envelope is an isosceles triangle, symmetric about $\rho = 0$. (See Figure 1.) The three points that define the triangle are $(0, (n_1n_2)^{-1})$, $(\rho_{max}, 0)$, and $(\rho_{min}, 0)$. The points on the parametric curve, i.e, the points between $(\rho_{min}, 0)$ and $(0, (n_1n_2)^{-1})$ and between $(0, (n_1n_2)^{-1})$ and $(\rho_{max}, 0)$, correspond to conventional mixtures (1).

We also assume that $n_1 \geq 3$ and $q = n_2/n_1$ is integer. In this case, Reilly (1991) and Peterson and Reilly (1991) characterize the minimum- and maximum-correlation pmfs for (X, Y):

$$g_{min}(x,y) = \begin{cases} \frac{1}{n_2}, & \text{if } q(n_1 - x) < y \leq q(n_1 - x + 1); \\ 0, & \text{otherwise;} \end{cases}$$

and

$$g_{max}(x,y) = \begin{cases} \frac{1}{n_2}, & \text{if } q(x-1) < y \leq qx; \\ 0, & \text{otherwise.} \end{cases}$$

It follows that

$$\rho_{max} = q\left((n_1^2 - 1)/(n_2^2 - 1)\right)^{\frac{1}{2}}$$

and that $\rho_{min} = -\rho_{max}$.

For any point (ρ_0, θ_0) in the parametric envelope, there is a unique pmf that is a mixture of $f_1(x)f_2(y)$, $g_{min}(x,y)$, and $g_{max}(x,y)$ (Peterson and Reilly, 1991; Reilly, 1991). Specifically,

$$\begin{aligned} g_p(x,y) &= \lambda_0^p f_1(x)f_2(y) + \lambda_{min}^p g_{min}(x,y) + \\ &\quad \lambda_{max}^p g_{max}(x,y), \end{aligned} \tag{2}$$

where

$$\lambda_0^p = n_1n_2\theta_0,$$

$$\lambda_{min}^p = (1 - n_1n_2\theta_0 - \rho_0/\rho_{max})/2,$$

and

$$\lambda_{max}^p = (1 - n_1n_2\theta_0 + \rho_0/\rho_{max})/2.$$

We refer to mixtures of the form (2) as parametric mixtures.

Parametric mixtures (2) are a more general class of mixtures that includes all conventional mixtures (1). Note that it is very easy to generate values of (X, Y) using (2). Reilly (1991) generates 0-1 knapsack problems using (2), and Pollock (1992) uses (2) to generate weighted set covering problems.

5 INDEPENDENT SAMPLING

Suppose that we generate N observations of (X, Y) with X and Y independent. In this case, $R_{N|f_1f_2}$ is asymptotically normally distributed with mean $\rho_{f_1f_2} = 0$ and variance

$$\sigma_{R_{N|f_1f_2}}^2 = \frac{\sigma_{XY|f_1f_2}^2}{N\sigma_X^2\sigma_Y^2},$$

where

$$\sigma_{XY|f_1f_2}^2 = \mu_X\mu_Y(7n_1n_2 - n_1 - n_2 - 5)/36.$$

If we are interested in predicting how likely we are to observe sample correlations with absolute value γ or greater, we can calculate

$$\Pr\left(|R_{N|f_1f_2}| \geq \gamma\right) = 2\left(1 - \Phi\left(\frac{\gamma\sigma_X\sigma_Y\sqrt{N}}{\sigma_{XY|f_1f_2}}\right)\right), \tag{3}$$

Table 1: Example values of $\Pr(|R_{N|f_1f_2}| \geq \gamma)$

γ	N		
	100	1000	10000
0.01	0.970	0.907	0.711
0.05	0.853	0.559	0.064
0.10	0.711	0.242	0
0.20	0.459	0.019	0
0.30	0.267	0	0
0.40	0.139	0	0
0.50	0.064	0	0
0.60	0.026	0	0
0.70	0.010	0	0
0.80	0.003	0	0
0.90	0.001	0	0
0.95	0	0	0

Table 2: Example values of N_{max}

γ	α		
	0.50	0.10	0.01
0.01	33259	197802	485040
0.05	1330	7912	19401
0.10	332	1978	4850
0.20	83	494	1212
0.30	36	219	538
0.40	20	123	303
0.50	13	79	194
0.60	9	54	134
0.70	6	40	98
0.80	5	30	75
0.90	4	24	59
0.95	3	21	53
0.99	3	20	49

where Φ is the cdf for the standard normal random variable. Clearly, as N increases, the probability (3) decreases. If N is large, there is likely to be little variability in the sample correlations between observed values of X and Y.

Suppose that $n_1 = 25$ and $n_2 = 100$. Table 1 shows values of $\Pr(|R_{N|f_1f_2}| \geq \gamma)$ for various combinations of N and γ. It is interesting to note that for nearly three of every ten samples of size 10000, $R_{N|f_1f_2}$ is likely to be between -0.01 and +0.01. Furthermore, it would be rare indeed to find $|R_{N|f_1f_2}| \geq 0.30$ if $N \geq 1000$.

The probabilities (3) in Table 1 suggest that if we hope to see some specified level of correlation between observed values of X and Y, we must be prudent in choosing N. Define N_{max} to be the maximum value of N such that $\Pr(|R_{N|f_1f_2}| \geq \gamma) \geq \alpha$. It can be shown that

$$N_{max} = \left\lfloor \left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2 \frac{\sigma^2_{XY|f_1f_2}}{\gamma^2\sigma^2_X\sigma^2_Y} \right\rfloor .$$

Suppose again that $n_1 = 25$ and $n_2 = 100$. Table 2 shows values of N_{max} for various combinations of γ and α. When values of X and Y are generated independently, N should be quite small if we hope to observe near-extreme sample correlations.

In addition to choosing a value for N, we may also decide in advance on how many test problems to generate. Consider the following Bernoulli random variable:

$$B = \begin{cases} 1 & \text{if } |R_{N|f_1f_2}| \geq \gamma; \\ 0 & \text{otherwise.} \end{cases}$$

If we generate ℓ test problems of size N with the pmf $f_1(x)f_2(y)$, then

$$B^\ell = \sum_{i=1}^{\ell} B_i \sim \text{Binomial}(\ell, \Pr(|R_{N|f_1f_2}| \geq \gamma)).$$

Suppose one of the requirements for our experiment is to have

$$\Pr(B^\ell \geq k) \geq \psi. \qquad (4)$$

Then our experiment should include at least $\ell_{min}(k)$ test problems, where $\ell_{min}(k)$ is the smallest integer ℓ that satisfies (4). If $k = 1$, (4) becomes

$$\begin{aligned} \Pr(B^\ell \geq 1) &= 1-(1-\Pr(|R_{N|f_1f_2}| \geq \gamma))^\ell \\ &= 1-e^{(\ell)\ln(1-\Pr(|R_{N|f_1f_2}|\geq\gamma))} \\ &\geq \psi, \end{aligned}$$

and it follows that

$$\ell_{min}(1) = \left\lceil \frac{\ln(1-\psi)}{\ln(1-\Pr(|R_{N|f_1f_2}| \geq \gamma))} \right\rceil .$$

For example, let $n_1 = 25$, $n_2 = 100$, $N = 1000$, $\psi = 0.20$, and

$$B = \begin{cases} 1 & \text{if } |R_{1000|f_1f_2}| \geq 0.20; \\ 0 & \text{otherwise.} \end{cases}$$

If we require enough test problems so that

$$\Pr(B^\ell \geq 1) \geq 0.85,$$

then $\ell_{min}(1) = 98$. In other words, we would need to generate at least 98 test problems of size 1000 to have a 0.85 chance that, for at least one problem,

Table 3: Values of $\ell_{min}(1)$ with $\psi = 0.50$

	N		
γ	100	1000	10000
0.01	1	1	1
0.05	1	1	11
0.10	1	3	3196
0.20	2	36	
0.30	3	1539	
0.40	5	239595	
0.50	11	1.38×10^9	
0.60	26		
0.70	72		
0.80	225		
0.90	794		
0.95	1568		
0.99	2764		

$R_{1000|f_1f_2} \leq -0.20$ or $R_{1000|f_1f_2} \geq 0.20$. Table 3 shows values of $\ell_{min}(1)$ for various combinations of N and γ with $\psi = 0.50$. We see that $\ell_{min}(1)$ grows very quickly as γ or N increases. If we require $\Pr(B^\ell \geq 2) \geq 0.85$, we would need to generate $\ell_{min}(2) = 174$ problems.

Let A_b be the negative binomial random variable that represents the number of test problems that are generated until there are b problems such that $|R_{N|f_1f_2}| \geq \gamma$. Then

$$\mu_{A_b} = \frac{b(1 - \Pr(|R_{N|f_1f_2}| \geq \gamma))}{\Pr(|R_{N|f_1f_2}| \geq \gamma)}$$

and

$$\sigma^2_{A_b} = \frac{b(1 - \Pr(|R_{N|f_1f_2}| \geq \gamma))}{(\Pr(|R_{N|f_1f_2}| \geq \gamma))^2}.$$

For our example, $\mu_{A_1} = 50.74$ and $\sigma^2_{A_1} = 2624.85$.

6 SAMPLING WITH INDUCED CORRELATION

Suppose that we wish to generate N observations of (X, Y) with $g_p(x, y)$. Then

$$\mu_{XY|g_p} = \lambda^p_0 \mu_X \mu_Y + \lambda^p_{min} \mu_{XY|g_{min}} + \lambda^p_{max} \mu_{XY|g_{max}},$$

$$\begin{aligned}\mu_{(XY)^2|g_p} = {} & \lambda^p_0 \mu_{(XY)^2|f_1f_2} + \lambda^p_{min} \mu_{(XY)^2|g_{min}} + \\ & \lambda^p_{max} \mu_{(XY)^2|g_{max}},\end{aligned}$$

and

$$\sigma^2_{XY|g_p} = \mu_{(XY)^2|g_p} - \mu^2_{XY|g_p}.$$

Peterson and Reilly (1991) and Reilly (1991) show that

$$\mu_{XY|g_{min}} = \mu_X(2n_2 + q + 3)/6$$

and

$$\mu_{XY|g_{max}} = \mu_X(4n_2 - q + 3)/6.$$

It can also be shown that

$$\mu_{(XY)^2|f_1f_2} = \mu_X \mu_Y (2n_1 + 1)(2n_2 + 1)/9,$$

$$\begin{aligned}\mu_{(XY)^2|g_{min}} = {} & \mu_X(6n_1n_2^2 + 4q^2 + 11qn_2 \\ & +9n_2^2 + 15n_1n_2 + 15n_2 + 15q \\ & +10n_1 + 5)/90,\end{aligned}$$

and

$$\begin{aligned}\mu_{(XY)^2|g_{max}} = {} & \mu_X(36n_1n_2^2 + 4q^2 - 19qn_2 \\ & +9n_2^2 + 45n_1n_2 + 15n_2 - 15q \\ & +10n_1 + 5)/90.\end{aligned}$$

Suppose that $n_1 = 25$ and $n_2 = 100$. In this case, $\rho_{max} = 0.99925$. Table 4, Table 5, and Table 6 show weights for mixing, values of $\mu_{XY|g_p}$ and $\sigma^2_{XY|g_p}$, and values of

$$\Pr(|R_{N|g_p} - \rho_0| \leq \eta) = 2\Phi\left(\frac{\eta\sigma_X\sigma_Y\sqrt{N}}{\sigma_{XY|g_p}}\right) - 1 \quad (5)$$

with $\eta = 0.05$ for various combinations of ρ_0, θ_0, and N, respectively. In Table 5 we see that there is a direct relationship between ρ_0 and $\sigma^2_{XY|g_p}$, and there is an indirect relationship between θ_0 and $\sigma^2_{XY|g_p}$ for a fixed ρ_0. We see in (5) and in Table 6 that $\sigma^2_{XY|g_p}$ affects how closely $R_{N|g_p}$ approximates ρ_0. The probabilities in Table 6 would be greater if we had used a larger tolerance on $|R_{N|g_p} - \rho_0|$, that is, if $\eta > 0.05$. See Table 7 for values of $\Pr(|R_{N|g_p} - \rho_0| \leq 0.10)$.

If we require that

$$\Pr(|R_{N|g_p} - \rho_0| \leq \eta) \geq \beta,$$

our test problem size N should be at least N_{min} where

$$N_{min} = \left\lceil (\Phi^{-1}((\beta + 1)/2))^2 \frac{\sigma^2_{XY|g_p}}{\eta^2\sigma^2_X\sigma^2_Y} \right\rceil.$$

Table 8 shows values of N_{min} for specified requirements of the form

$$\Pr(|R_{N|g_p} - \rho_0| \leq 0.05) \geq \beta$$

for various combinations of ρ_0, θ_0, and β. There is an indirect relationship between the entries in Table 8 and the corresponding entries in Tables 6 and 7. We

Table 4: Weights for example mixtures

ρ_0	θ_0	λ_0^p	λ_{max}^p	λ_{min}^p
$\rho_{max}/2$	0.0002	0.5	0.5	0
$\rho_{max}/2$	0.0001	0.25	0.625	0.125
$\rho_{max}/2$	0	0	0.75	0.25
0	0.0004	1	0	0
0	0.0003	0.75	0.125	0.125
0	0.0002	0.5	0.25	0.25
0	0.0001	0.25	0.375	0.375
0	0	0	0.5	0.5
$\rho_{min}/2$	0.0002	0.5	0	0.5
$\rho_{min}/2$	0.0001	0.25	0.125	0.625
$\rho_{min}/2$	0	0	0.25	0.75

Table 5: Values of $\mu_{XY|g_p}$, $\sigma_{XY|g_p}^2$

ρ_0	θ_0	$\mu_{XY|g_p}$	$\sigma_{XY|g_p}^2$
$\rho_{max}/2$	0.0002	760.5	459719.65
$\rho_{max}/2$	0.0001	760.5	468330.85
$\rho_{max}/2$	0	760.5	476942.05
0	0.0004	656.5	316761.25
0	0.0003	656.5	325372.45
0	0.0002	656.5	333983.65
0	0.0001	656.5	342594.85
0	0	656.5	351206.05
$\rho_{min}/2$	0.0002	552.5	186615.65
$\rho_{min}/2$	0.0001	552.5	195226.85
$\rho_{min}/2$	0	552.5	203838.05

Table 6: Values of $\Pr(|R_{N|g_p} - \rho_0| \leq 0.05)$

		N		
ρ_0	θ_0	100	1000	10000
$\rho_{max}/2$	0.0002	0.122	0.373	0.875
$\rho_{max}/2$	0.0001	0.121	0.369	0.872
$\rho_{max}/2$	0	0.120	0.366	0.868
0	0.0004	0.147	0.441	0.936
0	0.0003	0.145	0.436	0.932
0	0.0002	0.143	0.431	0.928
0	0.0001	0.141	0.426	0.925
0	0	0.139	0.421	0.921
$\rho_{min}/2$	0.0002	0.190	0.554	0.984
$\rho_{min}/2$	0.0001	0.186	0.544	0.982
$\rho_{min}/2$	0	0.182	0.534	0.979

Table 7: Values of $\Pr(|R_{N|g_p} - \rho_0| \leq 0.10)$

		N		
ρ_0	θ_0	100	1000	10000
$\rho_{max}/2$	0.0002	0.241	0.668	0.998
$\rho_{max}/2$	0.0001	0.239	0.664	0.998
$\rho_{max}/2$	0	0.237	0.659	0.997
0	0.0004	0.289	0.758	1.000
0	0.0003	0.285	0.751	1.000
0	0.0002	0.281	0.745	1.000
0	0.0001	0.278	0.739	1.000
0	0	0.275	0.733	1.000
$\rho_{min}/2$	0.0002	0.370	0.872	1.000
$\rho_{min}/2$	0.0001	0.362	0.864	1.000
$\rho_{min}/2$	0	0.355	0.855	1.000

Table 8: Values of N_{min} with $\eta = 0.05$

ρ_0	θ_0	β		
		0.50	0.80	0.95
$\rho_{max}/2$	0.0002	1931	6971	16304
$\rho_{max}/2$	0.0001	1967	7102	16610
$\rho_{max}/2$	0	2004	7232	16915
0	0.0004	1331	4804	11234
0	0.0003	1367	4934	11540
0	0.0002	1403	5065	11845
0	0.0001	1439	5195	12150
0	0	1476	5326	12456
$\rho_{min}/2$	0.0002	784	2830	6619
$\rho_{min}/2$	0.0001	820	2961	6924
$\rho_{min}/2$	0	857	3091	7230

see that N_{min} increases as ρ_0 or β increases. N_{min} would decrease if η is increased. For example, the entries in Table 8 would be reduced by a factor of 4 if $\eta = 0.10$.

Let

$$D = \begin{cases} 1 & \text{if } |R_{N|g_p} - \rho_0| \le \eta; \\ 0 & \text{otherwise.} \end{cases}$$

If we generate t test problems of size N with the pmf $g_p(x, y)$,

$$D^t = \sum_{i=1}^{t} D_i \sim \text{Binomial}(t, \Pr(|R_{N|g_p} - \rho_0| \le \eta)).$$

If one of the requirements for our experiment is to have

$$\Pr(D^t \ge k) \ge \zeta, \qquad (6)$$

then we must generate at least $t_{min}(k)$ test problems, where $t_{min}(k)$ is the smallest integer t that satisfies (6). When $k = 1$, (6) becomes

$$1 - e^{(t)\ln(1-\Pr(|R_{N|g_p}-\rho_0|\le\eta))} \ge \zeta,$$

and it follows that

$$t_{min}(1) = \left\lceil \frac{\ln(1-\zeta)}{\ln(1 - \Pr(|R_{N|g_p} - \rho_0| \le \eta))} \right\rceil .$$

Suppose that $n_1 = 25$, $n_2 = 100$, and $\eta = 0.05$, and that we require enough test problems so that

$$\Pr(D^t \ge 1) \ge 0.90.$$

Table 9: Values of $t_{min}(1)$ with $\zeta = 0.90$

ρ_0	θ_0	N		
		100	1000	10000
$\rho_{max}/2$	0.0002	18	5	2
$\rho_{max}/2$	0.0001	18	5	2
$\rho_{max}/2$	0	19	6	1
0	0.0004	15	4	1
0	0.0003	15	5	1
0	0.0002	15	5	1
0	0.0001	16	5	1
0	0	16	5	1
$\rho_{min}/2$	0.0002	11	3	1
$\rho_{min}/2$	0.0001	12	3	1
$\rho_{min}/2$	0	12	4	1

Table 9 shows values of $t_{min}(1)$ for various combinations of N and g_p with $\zeta = 0.90$. We see that $t_{min}(1)$ decreases as N increases. $t_{min}(1)$ is also directly related to $\sigma^2_{XY|g_p}$

The number of test problems that are generated until there are d problems such that $|R_{N|g_p} - \rho_0| \le \eta$ is a negative binomial random variable, W_d, with mean

$$\mu_{W_d} = \frac{d(1 - \Pr(|R_{N|g_p} - \rho_0| \le \eta))}{\Pr(|R_{N|g_p} - \rho_0| \le \eta)}$$

and variance

$$\sigma^2_{W_d} = \frac{d(1 - \Pr(|R_{N|g_p} - \rho_0| \le \eta))}{(\Pr(|R_{N|g_p} - \rho_0| \le \eta))^2}.$$

If $n_1 = 25$ and $n_2 = 100$, the expected number of problems generated, using a parametric mixture (2) with $\rho_0 = \rho_{max}/2$ and $\theta_0 = 0.0002$, to get one problem with $|R_{1000|g_p} - \rho_0| \le 0.05$ is only $\mu_{W_1} = 1.68$. Also, $\sigma^2_{W_1} = 4.52$.

7 DISCUSSION

We have compared two different input models for synthetic optimization problems with two types of coefficients, or for generating samples of a bivariate discrete random variable, with specified marginal distributions. Our findings have profound implications for computational experiments.

Many empirical evaluations of solution methods are carried out on large synthetic problems that are generated under the assumption that all coefficient types are mutually independent. We have seen that, as the size of the test problems increases, we expect the range of sample correlations between the two types of coefficients to fall within a narrower interval centered around 0. Hence, the synthetic problems become more similar, at least in a statistical sense, as the size of the test problems increases.

With the assumptions that we have made here, all of the test problems that would be generated with independent sampling would be generated with the unique pmf that corresponds to the point $(0, (n_1 n_2)^{-1})$ in the parametric envelope (Figure 1), that is, $f_1(x)f_2(y)$.

Results from computational studies on 0-1 knapsack problems (Martello and Toth, 1979; Balas and Zemel, 1980; Moore, 1989) and weighted set covering problems (Moore, 1990, Pollock, 1992) suggest that the performance of branch-and-bound methods and heuristics can degrade significantly as the correlation between the parameters in the objective function and the constraint(s) increases. Computational experience on large synthetic problems with independently generated parameters may be indicative of the median performance of the solution methods only, with little insight into worst-case behavior. Although it is common to experiment with large problems, our analysis here suggests that we might learn more about solution method performance if we used substantially smaller test problems when the parameters are generated independently.

Synthetic problems with induced correlation among the coefficient types can be generated easily. We can generate large problems for any specified point (ρ_0, θ_0) in the parametric envelope that allow us to get a more complete understanding of the performance of the solution method(s) of interest.

REFERENCES

Balas, E. and C.H. Martin. 1980. Pivot and Complement - A Heuristic for 0-1 Programming. *Management Science* **26**(1): 86–96.

Balas, E. and E. Zemel. 1980. An Algorithm for Large Zero-One Knapsack Problems. *Operations Research* **28**(5): 1130–1154.

Greenberg, H.J. 1990. Computational Testing: Why, How, and How Much. *ORSA Journal on Computing* **2**(1): 94–97.

Hoffman, K.L. and R.H.F. Jackson. 1982. In Pursuit of a Methodology for Test Mathematical Programming Software. *Evaluating Mathematical Programming Techniques*, Lecture Notes in Economics and Mathematical Systems No. 199, ed. J.M. Mulvey, 177–199. Springer-Verlag, New York.

Jackson, R.H.F., P.T. Boggs, S.G. Nash, and S. Powell. 1991. Guidelines for Reporting Results of Computational Results, Report of the Ad Hoc Committee. *Mathematical Programming* **49**: 413–425.

John, T.C. 1989. Tradeoff Solutions in Single Machine Production Scheduling for Minimizing Flow Time and Maximum Penalty. *Computers and Operations Research* **16**(5): 471–479.

Loulou, R. and E. Michaelides. 1979. New Greedy Heuristics for the Multidimensional 0-1 Knapsack Problem. *Operations Research* **27**(6): 1101–1114.

Martello, S. and P. Toth. 1979. The 0-1 Knapsack Problem. *Combinatorial Optimization*, eds. N. Christofides, A. Mingozzi, and C. Sandi, 237–279. John Wiley and Sons, New York.

Moore, B.A. 1989. Correlated 0-1 Knapsack Problems. IND ENG 854 Course Project, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, OH.

Moore, B.A. 1990. The Effect of Correlation on the Performance of Exact and Heuristic Procedures for the Weighted Set Covering Problem. M.S. Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, OH.

Peterson, J.A. 1990. A Parametric Analysis of a Bottleneck Transportation Problem Applied to the Characterization of Correlated Discrete Bivariate Random Variables. MS Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, OH.

Peterson, J.A. and C.H. Reilly. 1991. Maximin Probability Mass Functions for Simulating Bivariate Random Variables. Working Paper 1991-009, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, OH.

Pollock, G.A. 1992. Evaluation of Solution Methods for Weighted Set Covering Problems Generated with Correlated Uniform Random Variables. Undergraduate Honors Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, OH.

Potts, C.N. and L.N. Van Wassenhove. 1988. Algorithms for Scheduling a Single Machine to Minimize the Weighted Number of Late Jobs. *Management Science* **34**(7): 843–858.

Reilly, C.H. 1991. Optimization Test Problems with Uniformly Distributed Coefficients. *Proceedings of the 1991 Winter Simulation Conference*, eds. B. L. Nelson, W. D. Kelton, G. M. Clark, 866–874. Institute of Electrical and Electronics Engineers, Phoenix, Arizona.

Schmeiser, B.W. and R. Lal. 1982. Bivariate Gamma Random Variables. *Operations Research* **30**(2): 355–374.

AUTHOR BIOGRAPHY

CHARLES H. REILLY is an Associate Professor and Vice Chair in the Department of Industrial and Systems Engineering at The Ohio State University. He earned a B.A. in mathematics and business administration at St. Norbert College in 1979 and M.S. and Ph.D. degrees at Purdue University in 1980 and 1983, respectively. His current research interests are in the areas of empirical evaluation of solution methods for optimization problems and applications of optimization in manufacturing and telecommunications. Dr. Reilly is a member of ORSA, TIMS, IIE, and ASEE and the Mathematical Programming Interest Group Chair for the Operations Research Division of IIE.