Variance Reduction via Adaptive Control Variates (ACV)

Jae Yeong Lee

Abstract

Control Variate (CV) is very useful technique for variance reduction in a wide class of queueing network simulations. However, the loss in variance reduction caused by the estimation of the optimum control coefficients is an increasing function of the number of control variables. Therefore, in some situations, it is required to select an optimal set of control variables to maximize the variance reduction. In this paper, we develop the Adaptive Control Variates (ACV) method which selects an optimal set of control variates during the simulation adatively. ACV is useful to maximize the simulation efficiency when we need iterated simulations to find an optimal solution. One such an example is the Simulated Annealing (SA) because, in SA algorithm, we have to repeat in calculating the objective function values at each temperature. The ACV can also be applied to the queueing network optimization problems to find an optimal input parameters (such as service rates) to maximize the throughput rate with a certain cost constraint.

Key Words: Adaptive Control Variate, Variance Reduction, Simulation Optimization, Simulated Annealing

1. Introduction

Even though simulation is the most widely used technique of operations research in building and running detailed models of complex, real-world systems, it is in general computationally intensive. Therefore, numerous variance reduction techniques (VRTs) have been developed to improve the efficiency of simulation experiments. A comprehensive survey of VRTs is given in Wilson (1984) and Nelson (1987).

We are particularly interested in one VRT, called control variates (CVs). The basic idea for the method of CV is to take advantage of correlation between specified CVs and responses of a stochastic simulation. Suppose we can identify a $1 \times q$ vector of concomitant random variables $\boldsymbol{C} = (C_1,\cdots,C_q)$ that are generated by the simulation and that have a known, finite expectation $\mu c \equiv \mathrm{E[C]}$ as well as a strong correlation with the response variable Y whose mean $\theta \equiv \mu_Y \equiv \mathrm{E[Y]}$ is to be estimated. When using the method of CV, the unknown deviation $Y - E[Y] = Y - \theta$ is predicted as a linear combination of the known deviation $\boldsymbol{C} - \mu_C$. Thus we have the controlled response

$$Y_{CV} \equiv Y - \beta(C - \mu_C)', \qquad (1)$$

* Jae-Yeong Lee, P.O.Box 78-5, BCTP, YOOSUNG-KU CHUMOK-DONG 567, TAEJEON CITY, 305-153 Tel : (042) 870-1717, Fax : (042) 863-7568, e-mail : jlee7172 @kornet.mn.kr.

Where β is a $1 \times q$ vector of control coefficients and the notation A′ means the transpose of a row vector **A**. Since β is constant, Y_{CV} is an unbiased estimator of θ. Let $\sigma_{YC} \equiv Cov(Y,C)$ denote the $1 \times q$ vector of covariances $[Cov(Y,C_1),\cdots,Cov(Y,C_q)]$ and let $\Sigma_C \equiv Var(C)$ denote the q × q variance-covariance matrix of **C** , where we assume that Σ_C is positive definite. Lavenberg et al. (1982) showed that the variance of Y_{CV} is minimized by the optimal control coefficient vector

$$\beta^* = \sigma_{YC}\Sigma_C^{-1}. \qquad (2)$$

If σ_{YC} is not known, we have to use a sample estimator of β^* as follows :

$$\hat{\beta} = S_{YC}S_C^{-1}. \qquad (3)$$

where S_{YC} and S_C are sample estimators of σ_{YC} and Σ_C, respectively. We also denote S_C^{-1} at the inverse matrix of S_C.

The CV estimator of θ based on the random sample $\{(Y_i,C_i) : i=1,\cdots,n\}$ of size n is then defined as

$$\hat{\theta}_{CV}(n) \equiv \bar{Y} - \hat{\beta}\,(\bar{C} - \mu_C)\,' \;, \qquad (4)$$

where $\bar{Y}$ and $\bar{C}$ are the sample means of the $\{Y_i\}$ and $\{C_i\}$, respectively. If each vector (Y_i, C_i) has a joint multivariate normal distribution and $n \rangle q+2$, then $\hat{\theta}_{CV}(n)$ is an unbiased estimator of θ with variance

$$Var[\,\hat{\theta}_{CV}(n)] = \frac{n-2}{n-q-2}(1 - R^2_{YC})\frac{\sigma^2_Y}{n}, \qquad (5)$$

where $R^2_{YC} = \sigma_{YC}\Sigma_C^{-1}\,\sigma'_{YC}/\,\sigma^2_Y$ is the squared coefficient of multiple correlation between **Y** and **C**(Lavenberg et al. 1982). The product of the last two factors on the right hand side of (5) is the minimum variance obtainable if β were known. The term $(1 - R^2_{YC})$ is often called the *minimum variance ratio*. The factor $(n-2)/(n-q-2)$ represents the amount by which the variance is increased when β is unknown and is estimated by the method of least squares. This is called the *loss factor*. Equation (5) yields the *net variance ratio*, which is the ratio of the variances of the controlled and uncontrolled estimators of the mean response.

Since the *loss factor* is an increasing function of q, we have to find an optimal value of q to prevent inflating the variance. More specifically, the following issues arise in seeking to minimize the variance of the controlled simulation response:

(a) What is the optimal size (dimension) q of the vector of CVs?

(b) Which CVs should be in the best subset of all candidate CVs?

Currently, there are no general answers of these questions. However, many authors have contributed to provide some analytical solutions in simple cases and some rigorous experimental results in more complex situations. Question (a) has been addressed by Rubinstein and Marcus (1985) and Porta Nova and Wilson (1993). Question (b) has been discussed in the simulation literature by Lavenberg et al. (1982), Nozari et al. (1984), Añonuevo and Nelson (1988), and Bauer and Wilson (1992). Their ideas are described in Section 2.

In this paper we propose a method of adaptive control variates (ACVs) that selects an optimal set of control variates adaptively during iterated runs of a simulation model. Using the procedure of ACV, we develop an algorithm to handle both questions (a) and (b) simultaneously. The simulated annealing (SA) algorithm provides an excellent domain for applying the proposed ACV procedure since SA requires iterated simulations to seek optimal input parameters to minimize the expected value of a selected response function. In this case, effective VRTs are critical to reduce the large number of simulation replications generally required by the annealing procedure.

This paper is organized as follows. In Section 2, we discuss about the control-variate selection problem. In Section 3, as a building block of the ACV we introduce work variables (WVs) and explain why we select WVs among other control variables. In Section 4, the ACV procedure is formulated and we evaluate the ACV method by doing some experiments. Finally, in Section 5, our conclusions are addressed.

2. The Control-Variate Selection Problem

2.1 Optimal Number of Controls

In general, finding the optimal number of control variates to minimize the net variance ratio is difficult to implement analytically. Rubinstein and Marcus (1985) and Porta Nova and Wilson (1993) considered the situation in which a p-dimensional response **Y** vector and a q-dimensional control vector **C** jointly have a covariance matrix of the form

$$\Sigma = \left[\begin{array}{cccc|cccc} 1 & \gamma & \cdots & \gamma & \gamma & \gamma & \cdots & \gamma \\ \gamma & 1 & \cdots & \gamma & \gamma & \gamma & \cdots & \gamma \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma & \gamma & \cdots & 1 & \gamma & \gamma & \cdots & \gamma \\ \hline \gamma & \gamma & \cdots & \gamma & 1 & \gamma & \cdots & \gamma \\ \gamma & \gamma & \cdots & \gamma & \gamma & 1 & \cdots & \gamma \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma & \gamma & \cdots & \gamma & \gamma & \gamma & \cdots & 1 \end{array}\right] = \left[\begin{array}{c|c} \Sigma_Y & \Sigma_{YC} \\ \hline \Sigma_{YC} & \Sigma_C \end{array}\right]. \qquad (6)$$

For the scalar-response case $(p=1)$, Rubinstein and Marcus derive analytically the optimal size of the vector of control variates. For the multiresponse case $(p>1)$, the numerical results of Rubinstein and Marcus show that substantial variance reduction is achieved when the number of control variates is relatively small (approximately of the same order as the number of unknown parameters).

Porta Nova and Wilson (1993) also discussed the optimal selection of control variates for simulation experiments in which the objective is estimation of a multiresponse *metamodel*-that is, a regression model of the mean response expressed in terms of design variables that are relevant to the system being analyzed. Based on the same assumption (6) about the form of the covariance matrix, they derived the net variance ratio as follows:

$$\eta = \left(\frac{n-m-1}{n-m-q-1}\right)^{mp} \left[1 - \frac{q\gamma}{(q-1)\gamma+1} \cdot \frac{p\gamma}{(p-1)\gamma+1}\right]^{m}, \qquad (7)$$

where m is the number of design variables and p is the number of response variables. Recall that n and q are the number of independent replications of each design point and the number of control variates, respectively. If both m and p are set equal to one, then we have

$$\eta = \left(\frac{n-2}{n-q-2}\right)\left[1 - \frac{q\gamma^2}{(q-1)\gamma+1}\right]. \qquad (8)$$

This is the case that was treated analytically by Rubinstein and Marcus (1985). Hence, given the values of n and γ, we can find the optimal size of q to minimize η.

However, in general, the covariance matrix does not satisfy (6) so that we cannot use (8) and have to rely on experimental results to select the optimal size of the control variates. Porta Nova and Wilson concluded, based on their analysis of the repeated-measures random-effects covariance structure (6) and of an autoregressive covariance structure that frequently arises in econometric and time series applications, that as a function of the number of selected controls, the efficiency of the controlled point estimator is often relatively insensitive in the neighborhood of the optimal number of controls.

2.2. Optimal Subset of CVs

When we have q available control variates, the total number of possible subsets of control variates is 2^q. Therefore, an exhaustive search method is not considered in general because of its complexity and inefficiency. Funival and Wilson (1974) applied a *branch and bound* algorithm to find the best subsets of independent variables in a regression model without examining all possible subsets.

The *forward stepwise regression* procedure and its variations are also used to select the best subsets of control variates by Lavenberg et al. (1982), Nozari et al. (1984), and Añonuevo and Nelson (1988). A limitation of the forward stepwise regression search approach is that it presumes there is a single "best" subset of independent variables and seeks to identify that subset. However there is often no unique "best" subset. Another limitation of the forward stepwise regression routine is that it sometimes arrives at an unreasonable "best" subset when the independent variables are very highly correlated. See pages 454-459 in Neter et al. (1990) for a detailed discussion of the forward stepwise regression procedure. Furthermore, for multiresponse simulations, Bauer and Wilson (1992) proposed control-variate selection criteria that minimize the mean-square confidence-region volume. However, their selection criteria have to be calculated for all possible subsets of CVs to identify the best one.

3. Introduction and Selection of Work Variables (WVs)

In a queueing network, the typical random variables that are related to the system sojourn time are service-time variables and multinomial routing variables. Lavenverg et al. (1982) defined *work variables* by combining (that is, taking the product of) these two types of random variables. They showed experimentally that the minimum variance ratio, $(1 - R^2_{YC})$, for a control vector composed of work variables was substantially smaller than the minimum variance ratio for a comparable control vector composed of service-time variables alone or routing variables alone (they called these latter controls "flow variables"); moreover, the minimum variance ratio for a control vector of work variables was approximately the same as for the comparable combined control vector of service-time and routing control variates. In their experiments, Bauer and Wilson (1993) also showed that using both standardized service-time and routing variables is superior to using only standardized service-time variables. However, no theoretical explanation has been given for why work variables tend to achieve larger increases in efficiency when estimating the mean response of a simulation in comparison to the efficiency increases obtained with other control variables.

In this section, we compare two candidate control variates, which are standardized *service-time variables* (STVs) and standardized *work variables* (WVs). Standardized STVs were introduced by Wilson and Pritsker (1984), and standardized *routing variables* (RVs) were proposed by Bauer and Wilson (1993). The standardized work variables proposed in this paper are fundamentally different from the (unstandardized) work variables originally proposed by Lavenberg et al. (1982) in the following way: for a large class of regenerative queueing networks, the proposed standardized work variables asymptotically possess a multivariate normal distribution with a known, nonsingular covariance matrix. Moreover, the standardized work variables defined for different service centers (nodes) of the queueing network are asymptotically independent. The (unstandardized) work variables of Lavenberg et al. (1982) do not have such asymptotic behavior. In this subsection we establish these key properties of our standardized work variables.

Suppose the service-time process at service center k in a queueing network with J service centers is the IID sequence $\{ U_l(k) : l \geq 1 \}$, where $\mu_k \equiv \mathrm{E}[U_1(k)]$ and $\sigma_k^2 \equiv \mathrm{Var}[U_l(k)]$ respectively denote the mean and variance of the service times sampled at service center k for $k = 1,\cdots,J$. Let $a(k,t)$ be the number of service times that are completed at center k in the period $[0,t]$. A standardized STV is then defined as

$$V_k(t) = [\,a(k,t)]^{-1/2} \sum_{l=1}^{a(k,t)} \frac{U_l(k) - \mu_k}{\sigma_k} \quad \text{for } k = 1,\cdots, J. \tag{9}$$

Given a regenerative queueing system in which the asymptotic sampling rate at service center k is

$$\alpha_k \equiv \lim_{t\to\infty} a(k,t)/t > 0 \text{ with probability one for } k = 1,\cdots, J,$$

Wilson and Pritsker (1984) showed that the vector $V(t) = [\,V_1(t),\cdots,V_J(t)\,]$ of standardized service-time variables converges to a J-variate normal distribution with mean vector $\mathbf{0}_J$ (the $J \times 1$ vector of zeros) and covariance matrix $\boldsymbol{I}_J$ (the $J \times J$ identity matrix),

$$V(t) \xrightarrow{D} N_J(\mathbf{0}_J, \boldsymbol{I}_J) \ \ as \ \ t \to \infty. \tag{10}$$

Similarly, a standardized RV at service center k can be defined for each destination service center m to which a customer can be routed probabilistically after departing service center k. In general, at service center k there are

v_k ($v_k \geq 1$) nonzero routing probabilities $\{p(k,m) : m \in D_k\}$ corresponding to the destinations of customers departing center k, where

$$\sum_{m \in D_k} p(k,m) = 1;$$

and for each of these destinations there is a standardized routing control of the form

$$R_{km}(t) = \sum_{l=1}^{a(k,t)} \frac{I_l(k,m) - p(k,m)}{\{a(k,t)[1 - p(k,m)]p(k,m)\}^{1/2}} \text{ for } m \in D_k, 1 \leq k \leq J,$$

where $I_l(k,m)$ is an indicator variable such that $I_l(k,m) = 1$ if the lth customer departing from service center k is routed to service center m, and $I_l(k,m) = 0$ otherwise. Notice that the routing control variable $R_{km}(t)$ is only well-defined at a service center for which $v_k \geq 2$; and in this case we define

$$\boldsymbol{D}_k^\circ \equiv \boldsymbol{D}_k - max\{m : p(k,m) \rangle 0\}$$

to be the set of $v_k - 1$ service centers in $\boldsymbol{D}_k$ obtained by arbitrarily deleting the largest service-center index in $\boldsymbol{D}_k$. If at service center k we define the vector of routing controls

$$\boldsymbol{R}_k(t) \equiv [R_{km}(t) : m \in \overset{\circ}{\boldsymbol{D}_k}],$$

then Bauer (1987) proved that $\mathbf{R}_k(t)$ has expected value $\mathbf{0}_{v_k}$-1 the $(v_k$-$1) \times 1$ null vector ; moreover, $\mathbf{R}_k(t)$ converges in distribution to a multivariate normal distribution

$$\mathbf{R}_k(t) \xrightarrow{D} N_{v_k}\text{-}1\,(\mathbf{0}_{v_k}\text{-}1, \Sigma_{\boldsymbol{R}(k)}) \text{ as } t \rightarrow \infty, \tag{11}$$

with (v_k-1) × (v_k-1) asymptotic nonsingular covariance matrix $\Sigma_{\boldsymbol{R}(k)}$ with (m,s) element

$$(\Sigma_{\boldsymbol{R}(k)})_{m,s} = \begin{cases} 1, & \text{if } m = s, \\ -\left\{ \dfrac{p(k,m)p(k,s)}{[1-p(k,m)][1-p(k,s)]} \right\}^{1/2}, & \text{if } m \neq s, \end{cases}$$

for $m,s \in \overset{\circ}{\boldsymbol{D}_k}$. By combining the basic ideas of standardized service-time controls and standardized routing controls, we can define the $v_k \times 1$ control vector

$$\boldsymbol{W}_k(t) \equiv [W_{km}(t) : m \in D_k]$$

of standardized work variables at service center k, provided $v_k \geq 1$, where

$$W_{km}(t) = \begin{cases} [a(k,t)]^{-1/2} \displaystyle\sum_{l=1}^{a(k,t)} \frac{U_l(k)I_l(k,m) - \mu_k P(k,m)}{\Theta(k,m)}, & \text{if } a(k,t) \rangle 0, \\ 0, & \text{if } a(k,t) = 0 \end{cases} \tag{12}$$

for $m \in \boldsymbol{D}_k$ and $1 \leq k \leq J$, , and

$$\begin{aligned} \Theta(k,m) &= \{\text{Var}[U_l(k)I_l(k,m)]\}^{1/2} \\ &= \{\text{E}[U_l^2(k)I_l^2(k,m)] - \text{E}^2[U_l(k)I_l(k,m)]\}^{1/2} \\ &= \{E[U_l^2(k)]E[I_l(k,m)] - \text{E}^2[U_l(k)]\text{E}^2[I_l(k,m)]\}^{1/2} \\ &= \{(\sigma_k^2 + \mu_k^2)p(k,m) - \mu_k^2 p^2(k,m)\}^{1/2} \\ &= p^{1/2}(k,m)\{\sigma_k^2 + \mu_k^2[1-p(k,m)]\}^{1/2} \end{aligned}$$

The formulation of $\Theta(k,m)$ follows from the assumption that the lth service time, $U_l(k)$, and the associated routing indicator, $I_l(k,m)$, are sampled independently of each other and of all service times and routing indicators for previous customers.

For a regenerative queueing network with J nodes (service centers), the following theorem establishes the asymptotic behavior of the standardized work variables having the form (12).

Theorem 1. If (i) for each service center k with $v_k \geq 1$, the service time $U_l(k)$ has a probability density $f_{U(k)}(u)$ for all $u \in \Re^1$ ($1 \leq k \leq J$) ; and (ii) each service center k with $v_k \geq 1$ has a nonzero asymptotic throughput rate so that

$$\lim_{t \to \infty} a(k,t)/t = \alpha_k \rangle 0 \text{ with probability } 1$$

($1 \leq k \leq J$), then for each service center k with $v_k \geq 1$, we have

$$\boldsymbol{W}_k(t) \xrightarrow{D} N_{vk}(\mathbf{0}_{vk}, \Sigma_{W(k)}) \text{ as } t \rightarrow \infty,$$

where the (m,s) element of $\Sigma_{W(k)}$ is given by

$$(\Sigma_{W(k)})_{m,s} = \begin{cases} 1, & \text{if } m=s, \\ -\dfrac{\mu_k^2 p(k,m)p(k,s)}{\Theta(k,m)\Theta(k,s)}, & \text{if } m \neq s \end{cases}$$

for all $m, s \in \boldsymbol{D}_k$. Moreover, if we define

$$\mathcal{B} \equiv \{k : 1 \le k \le J \text{ and } vk \ge 1\}$$

and if we take

$$v \equiv \sum_{k=1}^{J} v_k,$$

then the work vectors $\mathbf{W}(t)$ and $\mathbf{W}_h(t)$ defined for two different service centers k and h ($h,k \in \mathcal{B}$) are asymptotically independent so that the overall vector of work variables

$$\boldsymbol{W}(t) \equiv [\, \boldsymbol{W}_k(t) : k \in \mathcal{B}\,]$$

is asymptotically normal

$$\boldsymbol{W}(t) \xrightarrow{D} N_v(\boldsymbol{O}_v, \Sigma_W) \text{ as } t \to \infty,$$

with the block-diagonal covariance matrix

$$\Sigma_W = diag\,[\, \Sigma_{W(k)} : k \in \mathcal{B}\,]$$

Proof. See Lee(1995). □

Using multiple linear regression in matrix terms ati ith simulation run, we assume that have

$$Y^{(i)} = \mu_Y + \beta_V(V^{(i)} - \mu_V)' + \varepsilon_V^{(i)}, \text{ where } \varepsilon_V^{(i)} \sim N(0, \sigma_V^2) \qquad (13)$$

so that the corresponding controlled response is $Y_V^{(i)} = Y^{(i)} - \beta_V(V^{(i)} - \mu_V)'$; and similarly we can take

$$Y^{(i)} = \mu_Y + \beta_W(\boldsymbol{W}^{(i)} - \mu_W)' + \varepsilon_W^{(i)}, \quad \text{where } \varepsilon_W^{(i)} \sim \mathbf{N}\,(\mathbf{0}, \sigma_W^2) \qquad (14)$$

so that the corresponding controlled response is $Y_W^{(i)} = Y^{(i)} - \beta_W(\boldsymbol{W}^{(i)} - \mu_W)'$, where $\varepsilon_V^{(i)}$ and $\varepsilon_W^{(i)}$ are residual vectors whose probability distributions are assumed to be normal with zero means. Let σ_V^2 and σ_W^2 be the variances of $\varepsilon_V^{(i)}$ and $\varepsilon_W^{(i)}$, respectively, for all $i = n_0, n_0+1, \cdots, n$ where $n_0 > q+2$ (q is number of controls used). Since $\mathrm{E}[\overline{Y}_V] = \mathrm{E}[\overline{Y}_W] = \mu_Y = \mathrm{E}[\overline{Y}]$, both $\overline{Y}_V$ and $\overline{Y}_W$ are unbiased estimators of μ_y, where

$$\overline{Y}_V = \sum_{i=1}^{n} Y_V^{(i)} / n, \quad \overline{Y}_W = \sum_{i=1}^{n} Y_W^{(i)} / n, \quad \overline{Y} = \sum_{i=1}^{n} Y^{(i)} / n,$$

Since

$$\mathrm{Var}[Y_V] = \sigma_V^2 = (1-R_{YV}^2)\sigma_Y^2 \text{ where } R_{YV}^2 = \sigma_{YV}\Sigma_V^{-1}\sigma'_{YV}/\sigma_Y^2,$$

$$\mathrm{Var}[Y_W] = \sigma_W^2 = (1-R_{YW}^2)\sigma_Y^2 \text{ where } R_{YW}^2 = \sigma_{YW}\Sigma_W^{-1}\sigma'_{YW}/\sigma_Y^2,$$

it is well known that $\mathrm{Var}[Y_V] \le \sigma_Y^2$ and $Var[Y_W] \le \sigma_Y^2$.

However, $\mathrm{Var}[Y_V]$ and $\mathrm{Var}[Y_W]$ are difficult to compare analytically. We claim that WV is more effective than STV in terms of variance reduction of the simulation response estimation. In order to prove this claim, we have to show either

$$\mathrm{Var}[Y_V] \ge \mathrm{Var}[Y_W] \qquad (15)$$

$$\frac{\mathrm{Var}[Y_V]}{\mathrm{Var}[Y_W]} \ge 1. \qquad (16)$$

Relation (16) can also be rewritten as

$$\frac{R_{YV}^2}{R_{YW}^2} = \frac{\beta_V\sigma'_{YV}}{\beta_W\sigma'_{YW}} \le 1. \qquad (17)$$

However, it is not possible to show (17) without knowing exact values of the control coefficient vectors β_V, β_W and the covariance vectors σ_{YV}, σ_{YW}. Therefore, in order to check experimentally the significance of the difference in effectiveness of the two control vectors $\mathbf{V}$ and $\mathbf{W}$, first we perform a hypothesis test comparing R_{YV}^2 and R_{YW}^2 ; and then we perform simulation experiments involving both control to provide some numerical evidence to substantiate the claim (17).

3.1 Hypothesis Test for Multiple Correlation Coefficients

For the hypothesis test and the experiments, we considered a simulation model of the machine repair system shown in Figure 1. This queueing network could be used as a building block for more complex models of production lines, machine repair systems, multiprogrammed computer systems, etc. Center 1 is the main operating center for the machines, which are supposed to break down independently with rate λ for each machine. Initially, there are N operating machines and M spare machines. Whenever a machine breaks down, it is sent immediately to one of the service centers $2, 3, \cdots, J$ based on the type of machine failure, where a failure of type m occurs with probability $p(1,m)$ for $m = 2, \cdots, J$. After completion of the required service at the selected center k with rate $1/\mu_k$, the repaired machine will be sent back to the main operating center where it will either go back into operation or it will go into the pool of spares. It is obvious that this type of queueing network has the regenerative property and that the hypotheses of Theorem 1 are satisfied.

To do the hypothesis test for comparing R^2_{YV} and R^2_{YW}, we considered two types of control vectors $\boldsymbol{V} = (V_2, \cdots, V_J)$ and $\boldsymbol{W} = (W_{1,2}, \cdots, W_{1,J})$. From each simulation run, both STV and WV data sets are sampled using the following formulas:

$$V_k(t) = [a(k,t)]^{-1/2} \sum_{l=1}^{a(k,t)} \frac{U_l(k) - \mu_k}{\sigma_k}$$
$$= \frac{[a(k,t)]^{1/2}[\overline{U}(k) - \mu_k]}{\sigma_k} \text{ for } k = 2, \cdots, J,$$

and

$$W_{km}(t) = [a(k,t)]^{-1/2} \sum_{l=1}^{a(k,t)} \frac{U_l(k) I_l(k,m) - \mu_k p(k,m)}{\Theta(k,m)}$$
$$= \frac{[a(k,t)]^{1/2}[(1/a(k,t)) \Sigma_{l=1}^{a(k,t)} U_l(k) I_l(k,m) - \mu_k p(k,m)]}{p^{1/2}(k,m)\{\sigma_m^2 + \mu_m^2[1-p(k,m)]\}^{1/2}}$$

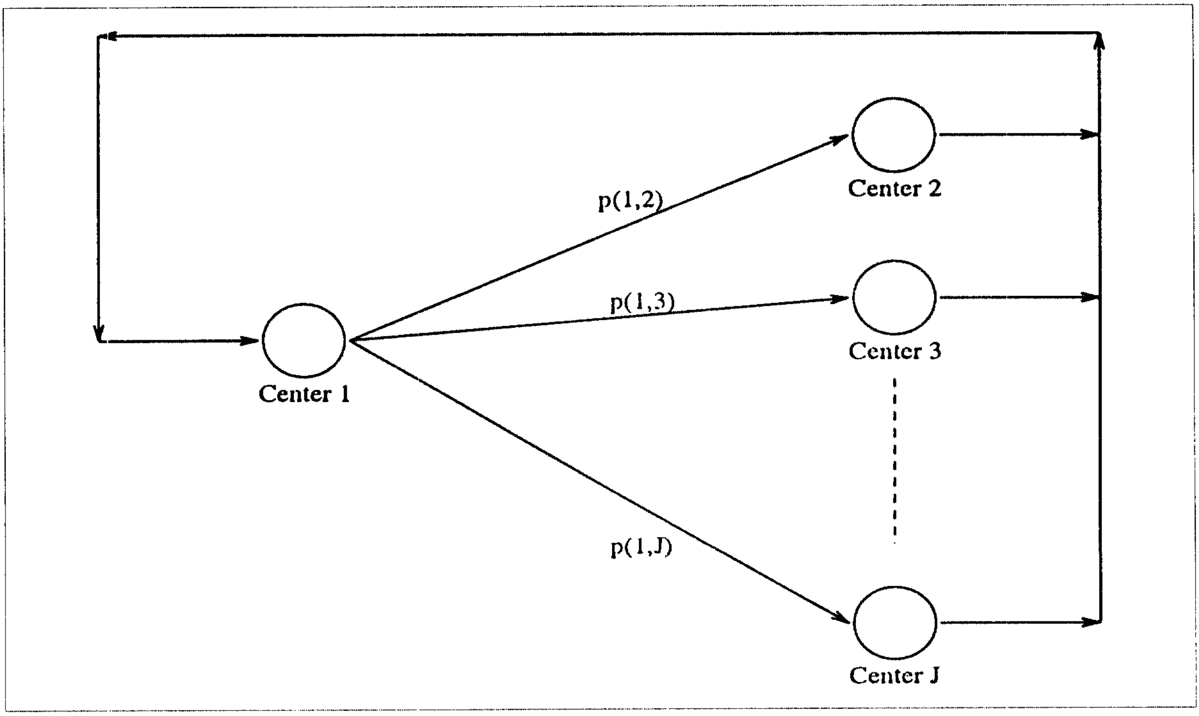

Figure 1 : Closed queueing network of machine repair system

for $2 \le m \le J$ and $k = 1$.

Let q be the number of control variates used (here, $q = J - 1$). Input data used for sampling are $J = 9$ $(q = 8)$ service centers, $N = 20$ machines, $M = 5$ spares, and the branching probabilities to centers $2, 3, \cdots, 9$ are given respectively by

$$\left.\begin{array}{l} p(1,2)=0.27,\ p(1,3)=0.18,\ p(1,4)=0.15,\ p(1,5)=0.12, \\ p(1,6)=0.10,\ p(1,7)=0.07,\ p(1,8)=0.06,\ p(1,9)=0.05 \end{array}\right\}$$

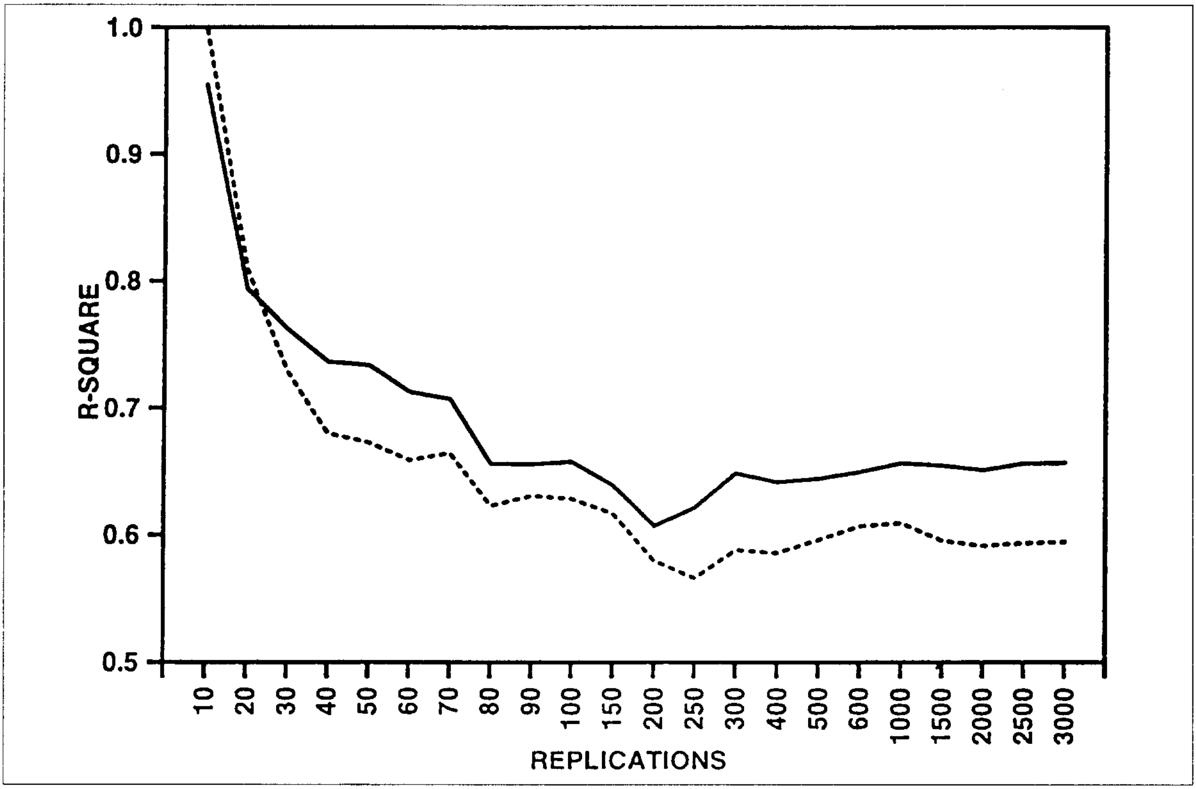

Figure 2 : Selection of stabilization point for $\hat{R}^2_{YV}$ and $\hat{R}^2_{YW}$.

The rate λ for each machine breakdown is 1.0 so that the time to failure for each machine is randomly sampled from an exponential distribution with mean μ_1=1.0 time units; and the repair time at service centers $2,\cdots,J$ are randomly sampled from an exponential distribution with mean μ_m = 0.2 time units for m=2,3,$\cdots$,J. The simulation ending time of each run is 1,000 time units, and the data collected for one fourth of that time (that is, the first 250 time units) is cleared to remove the initial-condition bias. The time units in this experiment are hours.

Next, some finite number of simulaton runs, n^*, is selected at the point where a stable esimate of R^2_{YC} is reached. To do this, we used the SAS procedure "**reg**" to compute estimates R^2_{YC} for different numbers of runs. Figure 2 reveals that a stable estimate of R^2_{YC} is obtained with n^* = 300 independent runs. This is insured by the p-values because the p-values of all regression coefficient estimators $\hat{\beta}_m$, for $m = 2, \cdots, J$, are less than 10^{-4} for both **V** and **W** when n^* = 300. This means that all control variates significantly contributed to both regression models when n^* = 300. Thus, 10 independent sets of 300 simulation runs are sampled and regressed; and the resulting 10 estimates of R^2_{YV} and R^2_{YW} are displayed in Table 1.

Therefore, the total number of simulation runs is 3,000 (= 10× 300). From Table 1, the t-statistic for the average of the differences $\hat{R}^2_{YW}$-$\hat{R}^2_{YV}$ is 8.835 ($=\sqrt{10}\times 0.06172/0.02209$), whose p-value is less than 0.0005, which is significant at the level $\alpha < 0.0005$ for 9 degrees of freedom. This shows that **W** yields a larger multiple correlation coefficient than **V** in the given stochastic system.

Table 1. 10 independent estimates $\hat{R}^2_{YV}$ and $\hat{R}^2_{YW}$

Est. #	$\hat{R}^2_{YV}$	$\hat{R}^2_{YW}$	$\hat{R}^2_{YW}$ - $\hat{R}^2_{YV}$
1	0.5879	0.6481	0.0602
2	0.6287	0.6633	0.0346
3	0.6400	0.6815	0.0415
4	0.5601	0.6727	0.1126
5	0.6022	0.6715	0.0693
6	0.5789	0.6347	0.0558
7	0.6239	0.6747	0.0508
8	0.5838	0.6719	0.0881
9	0.6303	0.6772	0.0469
10	0.6156	0.6730	0.0574
Mean	0.60514	0.6686	0.06172
Std.Dev.	0.02519	0.01379	0.02209

3.2 Experimental Results for the Controls V and W

In order to compare the actual efficiency of **V** and **W** as control vectors, we conducted a macroexperiment that consists of m microexperiments; and each microexperiment consists of n independent replications of the simulation. Let $Y_{i,j}$ be a response of interest from the ith replication in the jth microexperiment ($i = 1, \cdots, n$ and $j = 1,\cdots,m$), and let $C_{i,j}$ denote the corresponding vector of control variates such that $C_{i,j} = [C_{i,j,1},\cdots,C_{i,j,q}]$, where q is the total number of controls used. Let $V = [V_{i,j,1},\cdots,V_{i,j,q}]$ be sample mean vector of service-time control variables accumulated over the simulation time t, and let $W = [W_{i,j,1},\cdots,W_{i,j,q}]$ be the corresponding work control vectors. Then, again, we consider two types of control variates as follows:

$$C_{i,j} = \begin{cases} V_{i,j}, & \text{when STV is used,} \\ \mathbf{W}_{i,j}, & \text{when WV is used.} \end{cases}$$

The variances of the point estimators were estimated as follows. On the jth microexperiment ($j = 1,\cdots,m$), we have the mean responses and estimates of control (regression) coefficients as follows:

$$\bar{Y}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{i,j},$$

$$\bar{C}_j = \frac{1}{n}\sum_{i=1}^{n} C_{i,j},$$

$$\hat{\beta}_j = S^{(j)}_{YC}\ [S^{(j)}_C]^{-1}$$

where

$$S^{(i)}_{YC} = (n-1)^{-1}\Sigma^n_{i=1}(Y_{i,j}-\bar{Y}_j)(C_{i,j}-\bar{C}_j), \tag{18}$$

$$S^{(j)}_C = (n-1)^{-1}\Sigma^n_{i=1}(C_{i,j}-\bar{C}_j)'(C_{i,j}-\bar{C}_j), \tag{19}$$

and

$$[S^{(j)}_Y]^2 = (n-1)^{-1}\Sigma^n_{i=1}(Y_{i,j}-\bar{Y}_j)^2 \tag{20}$$

for $j = 1,\cdots,m$. Then, the jth controlled mean response is

$$\bar{Y}_j(\hat{\beta}_j) = \bar{Y}_j - \hat{\beta}_j(\bar{C}_j - \mu_C)' \text{ for } j = 1,\cdots,m; \tag{21}$$

and the variance of the controlled mean response in the jth microexperiment is estimated by

$$\begin{aligned} \widehat{\mathrm{Var}}[\bar{Y}_j(\overline{\beta}_j)] &= \{\widehat{\mathrm{Var}}[\bar{\mathrm{Y}}_\mathrm{j}] - [\mathrm{S}^{(j)}_{\mathrm{YC}}][\mathrm{S}^{(j)}_{\mathrm{C}}]^{-1}[\mathrm{S}^{(j)}_{\mathrm{YC}}]'\}\frac{\mathrm{n-2}}{\mathrm{n-q-2}} \\ &= \widehat{Var}[\bar{Y}_j](1-[\hat{R}^{(j)}_{YC}]^2)\frac{n-2}{n-q-2} \\ &= \frac{[S^{(j)}_Y]^2}{n}(1-[\hat{R}^{(j)}_{YC}]^2)\frac{n-2}{n-q-2}. \end{aligned} \tag{22}$$

Then, an *internal estimator* of the variance of the controlled mean response is the average of the sample variances of the form (22) computed across the m microexperiments as follows:

$$\widehat{Var}_{(int)}[\overline{Y}(\hat{\beta}_C)] = \frac{1}{m}\sum_{j=1}^{m}\widehat{Var}[\overline{Y}_j(\hat{\beta}_j)].$$

However, without the normality assumption on the response and the controls, it has been found that the internal variance estimator tends to underestimate the true variance of the controlled mean response $\overline{Y}_j(\hat{\beta}_j)$. In other words,

$$E\{\widehat{Var}_{(int)}[\overline{Y}(\hat{\beta}_C)]\} \le Var[\overline{Y}_j(\hat{\beta}_j)],$$

with equality when the response and the controls have a joint normal distribution. This was also found in Lavenberg et al. (1982).

Therefore, we also consider an *external variance estimator* to compare the efficiency of the two candidates **V** and **W** for the control vector **C**:

$$\widehat{Var}_{(ext)}[\overline{Y}(\hat{\beta}_c)]= \frac{1}{m-1}\sum_{j=1}^{m}[\overline{Y}_j(\hat{\beta}_j)-\overline{\overline{Y}}(\hat{\beta})^2,$$

where

$$\overline{\overline{Y}}(\hat{\beta}) = \frac{1}{m}\sum_{j=1}^{m}\overline{Y}_j(\hat{\beta}_j).$$

The external variance estmator is based on a macroexperiment consisting of m IID microexperiments, each with n simulation runs; and this is unbiased estimator of $Var[\overline{Y}(\hat{\beta}_C)]$ such that

$$E\{\widehat{Var}_{(ext)}[\overline{Y}(\hat{\beta}_C)]\} = Var[\overline{Y}(\hat{\beta}_C)].$$

Now, we will compare the following two variance estimators of the controlled responses:

$$\widehat{Var}[\overline{Y}_{stv}] \equiv \widehat{Var}_{(ext)}[\overline{Y}(\hat{\beta}_V)] \text{ and}$$
$$\widehat{Var}[\overline{Y}_{WV}] \equiv \widehat{Var}_{(ext)}[\overline{Y}(\hat{\beta}_W)].$$

To do these experiments, the closed queueing network of Figure 1 is used. One hundred microexperiments are executed (i.e., m = 100), and the number of runs for each microexperiment, n, is in the range $15 \le n \le 30$. The rest of the input data are the same as in the hypothesis test for comparing R^2_{YV} and R^2_{YW} in the previous section. Note that the number of runs in each microexperiment, n, should be greater than or equal to 11 to prevent the loss factor from exploding. Table 2 shows the external variance estmators for different values of n. The second column represents the variances when no control variates are applied (that is, with crude simulation). The third and the fourth columns are external variance estimators based on the macroexperiments in which the control vectors **V** and **W** are respectively applied. The percentages of variance reduction for both **V** and **W** are also shown in the fifth and the sixth columns and are denoted by VR_{stv} and VR_{wv}, respectively. These experimental results clearly show that applying **W** is more efficient than using **V** in the given stochastic system.

Table 2. External variance estimators and % variance reduction for different n.

n	$\widehat{Var}[\overline{Y}]$	$\widehat{Var}[\overline{Y}_{stv}]$	$\widehat{Var}[\overline{Y}_{WV}]$	$VR_{stv}(\%)$	$VR_{WV}(\%)$
15	0.7235	0.6163	0.5217	15	28
18	0.5295	0.3629	0.3183	31	40
20	0.5392	0.4116	0.3093	24	43
23	0.4830	0.3553	0.3028	26	37
25	0.4653	0.2457	0.1919	47	59
28	0.4514	0.2069	0.1652	54	63
30	0.3635	0.2127	0.1417	41	61

Table 3. External and internal variance estmators for different n.

n	$\widehat{Var}_{(ext)}[\overline{Y}_{stv}]$ (%)	$\widehat{Var}_{(ext)}[\overline{Y}_{WV}]$ (%)	$\widehat{Var}_{(int)}[\overline{Y}_{stv}]$ (%)	$\widehat{Var}_{(int)}[\overline{Y}_{WV}]$ (%)
15	0.6163(15)	0.5217(28)	0.3433(53)	0.2935(59)
18	0.3629(31)	0.3183(40	0.2551(52)	0.2265(57)
20	0.4116(24)	0.3093(43)	0.2329(57)	0.1954(64)
23	0.3553(26)	0.3028(37)	0.1935(60)	0.1628(66)
25	0.2457(47)	0.1919(59)	0.1766(62)	0.1540(67)
28	0.2069(54)	0.1652(63)	0.1585(65)	0.1363(70)
30	0.2127(41)	0.1417(61)	0.1503(59)	0.1282(65)

In Table 3, both external (second and third columns) and internal (fourth and fifth columns) variance estimators for **V** and **W** are displayed. As we discussed earlier, internal variance estimators (IVEs) are smaller than the corresponding external variance estimators (EVEs) in all cases in this experiment. The percentages of variance reduction are also given in the parentheses for all cases. Again, the percentage of variance reduction for **W** is larger than that for **V** in all cases in Table 3. Note that as n increases, the IVEs decrease monotonically and more linearly than the EVEs do. One important observation from this experiment is that **W** has more stable variance estimators than **V** does (see the second column in Table 3－the EVEs for **V** are not monotonically decreasing as n increases). This fact is also shown in Table 1, where the sample squared multiple correlation coefficients, $\hat{R}^2_{YV}$ and $\hat{R}^2_{YW}$, are compared. In Table 1, the standard deviation of 10 independent estimates of $\hat{R}^2_{YW}$ is much smaller than that of $\hat{R}^2_{YV}$ (0.01379 〈 0.02519). The stability of the EVEs for **W** could be another advantage for using **W** as a control vector instead of using **V**.

Based on not only the hypothesis test for the multiple correlation coefficients but also on the experimental results, we showed that **W** is a better control vector than **V** in terms of the variance reduction of the controlled simulation responses. One intuitive reason for this is that the flow (routing) effects of the network are taken into account in **W**, whereas **V** only takes into account service-time effects. This flow effect is caused by the difference between the various branching probabilities from each service center. That is, the larger the differences that we have among the routing probabilities from a given service center, the larger the correlation (between control variates and responses of the simulation) that we obtain. In other words, if we have the same branching probability to each service center, then the relative effectiveness of using the control vector **W** will be reduced. This flow effect of the work variable has been explained in terms of the routing control variable in the television-set inspection example of Bauer and Wilson (1993). In their experimental results, the smaller the probability of branching to the adjustor, the larger the reduction in confidence interval length for the average system sojourn time.

4. Adaptive Control Variates (ACVs)

4.1 Adaptive Selection of Subsets of WVs

Finding an optimal subset of control variates is most important when a relatively small number of simulation runs are available, because the loss factor in (5) will tend to unity as the number of runs gets large. In the application of SA to stochastic combinatorial optimization problems (SCOPs), a small number of simulation runs is needed to speed up the algorithm. That is, the smaller the number of runs we have at each temperature while satisfying the required precision for estimation of the mean simulation response, the faster will be the performance of the SA algorithm. In this case, variance reduction in estimating the mean response of the simulation is required to reduce the number of simulation runs.

Let $\mathbf{Y}' = (Y_1,\cdots,Y_n)$ be the n-dimensional vector of independent observations, where $Y_i, i = 1,\cdots,n$, is obtained from the ith simulation run. Suppose each observation $(Y_i, C_i)'$ is normally distributed

$$\begin{bmatrix} Y_i \\ \boldsymbol{C}'_i \end{bmatrix} \sim N_{q+1}\left[\begin{bmatrix} \mu_Y \\ \boldsymbol{O}_q \end{bmatrix}, \begin{bmatrix} \sigma^2_Y & \sigma_{YC} \\ \sigma_{CY} & \Sigma_C \end{bmatrix}\right] \text{for } i = 1.\cdots,n \qquad (23)$$

where μ_Y is the unconditional expected value of Y_i Then, give the control vectors $\{\mathbf{C}_i : i=1,\cdots,n\}$, the conditional expected value of **Y**, E $[\mathbf{Y} | \boldsymbol{C}_i : i=1,\cdots,n]$, can be written as $X\delta'$, where **X** is the $n \times (q+1)$ matrix

$$X = \begin{bmatrix} 1 & C_{11} & \cdots & C_{1q} \\ 1 & C_{21} & \cdots & C_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & C_{n1} & \cdots & C_{nq} \end{bmatrix}$$

and $\delta = (\mu_Y, \beta_1, \cdots, \beta_q)$ is the $(q+1)$-dimensional vector of unknown control coefficients. Notice that if Σ_C, the covariance matrix of the control vector in (23) is nonsingular, then it follows from Proposition 1 of Porta Nova and Wilson (1993) that **X** has rank $q+1$ with probability one. Note also that both μ_V and μ_W are zero vectors in (13)−(14) because **V** and **W** are standardized control variate vectors. Furthermore, given $C_i=(C_{i1},\cdots,C_{iq})$ for $i = 1, 2, \cdots, n$, we can write

$$Y \mid C_i : i=1,\cdots,n \sim N_n[X\delta', \sigma_Y^2(1-R_{YC}^2)I_n], \qquad (24)$$

where $\mathbf{I}_n$ denotes the n-dimensional identity matrix.

Now, let σ_{YC} and Σ_C denote the covariance vector between Y and C and the covariance matrix of C, respectively:

$$\sigma_{YC} \equiv \mathrm{Cov}(Y_i C_i) = \mathrm{E}[(Y_i - \mu_y)(C_i - \mu_C)],$$

$$\Sigma_C \equiv \mathrm{Cov}(C_i) = \mathrm{E}[(C_i - \mu_C)'(C_i - \mu_C)].$$

Then, it follows directly from (24) that

$$Y_i \mid C_i \sim N(\mu_Y + \beta C'_i, \tau^2), \text{ for } i=1,\cdots,n, \qquad (25)$$

where

$$\beta = \sigma_{YC}\Sigma_C^{-1},$$

$$\tau^2 = \sigma_Y^2 - \sigma_{YC}\Sigma_C^{-1}\sigma'_{YC} = \sigma_Y^2(1-R_{YC}^2).$$

Since both σ_{YC} and Σ_C are frequently unknown in practice, β and τ^2 have to be estimated. In terms of the statistics

$$S_{YC} = (n-1)^{-1}\sum_{i=1}^{n}(Y_i-\bar{Y})(C_i-\bar{C}),$$

$$S_C = (n-1)^{-1}\sum_{i-1}^{n}(C_i-\bar{C})'(C_i-\bar{C}),$$

$$\hat{\sigma}_Y^2 = (n-1)^{-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2,$$

we have estimates of β and τ^2 as follows:

$$\hat{\beta} = S_{YC}S_C^{-1},$$

$$\hat{\tau}^2 = \hat{\sigma}_Y^2 - S_{YC}S_C^{-1}S'_{YC} = \hat{\sigma}_Y^2(1-\hat{R}_{YC}^2).$$

Under the assumption that each (Y_i,C_i) has a multivariate normal distribution, Nozari et al. (1984) showed that given $\{C_i : i=1,\cdots,n\}$, the conditional covariance matrix of the estimator $\hat{\beta}$ of the regression coefficient vector β is

$$\mathrm{Var}[\hat{\beta}_{CV} \mid C_i : i=1,\cdots,n]$$

$$= \frac{n-2}{n-q-2}\tau^2(X'X)^{-1} \text{ if } n-q-2 > 0. \qquad (26)$$

The forward stepwise regression procedure is the most widely used method to select the best subset of control variates even though it has some limitations. One obvious reason that we do not want to consider this method is its burden of computing. In this section, we develop a heuristic method using Student t-statistics to select a good subset of control variates by adding a small computational burden during the simulation.

Suppose we use q control variates from J service centers (i.e., $q \le J$). From (25), we have the conditional variance of each estimated control coefficient as follows:

$$\mathrm{Var}[\hat{\beta}_k \mid C_i : i=1,\cdots,n] = \sigma^2(\hat{\beta}_k)$$

$$= \frac{n-2}{n-q-2}\tau^2(X'X)^{-1}_{k+1,k+1} \qquad (27)$$

if $n-q-2 > 0$, where $(X'X)^{-1}_{k\ ,k}$ is the $(k+1)$th diagonal element of the inverse matrix $(X'X)^{-1}$ for $k = 0,\cdots,q$. Then, we have corresponding t-statistics

$$t_k = \frac{\hat{\beta}_k}{\hat{\sigma}(\hat{\beta}_k)} \, for \; k = 1,\cdots,q. \qquad (28)$$

Let $t_{(k)}$ be the t-statistic with the kth largest absolute value, and let t_{sum} be the sum of all q absolute t-statistics

$$t_{\text{sum}} = \sum_{k=1}^{q} | t_k |.$$

Let $\hat{R}^2_{\max}$ be the maximum estimated squared multiple correlation coefficient that is obtained when all q control variates are used (see Fact 2 below), and let c be some user-specified constant such that $0 \langle c \cdot \hat{R}^2_{\max} \langle 1$. Note that $c \cdot \hat{R}^2_{\max}$ is the minimum accumulated percentage of t_{sum} that must be achieved by adding additional controls to the selected subset of controls—that is, if $C_{(k)}$ is the control variate corresponding to the kth largest absolute t-statistic $| t_{(k)} |$ so that

$$| t_{(1)} | \geq | t_{(2)} | \geq \cdots \geq | t_{(q)} |,$$

then the size q^* of the selected control of CVs is the smallest k such that

$$\frac{\Sigma_{j=1}^{k} | t_{(j)} |}{t_{\text{sum}}} \geq \text{c} \cdot \hat{R}^2_{\max} ;$$

and the selected control vector is

$$C(q^*) \equiv [C_{(1)}, C_{(2)}, \cdots, C_{(q^*)}].$$

```
begin
  [1] Sort absolute $t$-statistics in decreasing order. Let $k$
      $\leftarrow 1$, $\psi_{sum} \leftarrow 0$.
      Compute $t_{sum} = \sum_{j=1}^{q} |t_{(j)}|$ and $\psi_j \leftarrow | t_{(j)} | / t_{sum}$
            $for\ j = 1, \cdots, q$.
  [2] Cumulate $\psi_k : \psi_{sum} \leftarrow \psi_{sum} + \psi_k$
  [3] if ($\psi_{sum} \rangle c \cdot R^2_{max}$) then
        $q \leftarrow k$ ;
        go to [4]
      else
        $k \leftarrow k + 1$ ;
        go to [2]
      end if
  [4] Take C(q) $\leftarrow [C_{(1)}, C_{(2)}, \cdots, C_{(q)}]$ and compute $\hat{R}^2_{YC(q)}$.

  [5] if(1- $\hat{R}^2_{YC(q)} \langle \frac{n-q-2}{n-q}$) then go to [6]
      else
        $q \leftarrow q-1$.
        go to [4]
      end if
  [6] Deliver $q^* \leftarrow q$ and $C \equiv C(q^*) \equiv [C_{(1)}, C_{(2)}, \cdots, C_{(q)}]$.
  end
```

Figure 3. Subset selection procedure of the CVs

An algorithmic statement of this subset selection procedure for control variates is in Figure 3.

The basic idea of this algorithm is based on the following fact:

Fact 2. If (**Y, C**) has a multivariate normal distribution, then

$$R^2_{Y(C_1,\cdots,C_k,C_{k+1})} \geq R^2_{Y(C_1,\cdots,C_k)} \text{ for } k = 1,\cdots,q\text{-}1. \tag{29}$$

That is, the multivariate correlation coefficient between Y and C, R_{YC}, increases as the number of control variates increases. □

Fact 2 follows immediately from Theorem 2.5.4 of Anderson (1984) and the definition of the multiple correlation coefficient given on page 40 of Anderson (1984).

Note that the algorithm in Figure 3 does not guarantee that $C(q^*)$ is the best subset which gives the largest variance reduction in simulation responses. However, if we use the procedure in Figure 3, the limitations of forward stepwise regression procedure can be avoided by speeding up the simulated annealing (SA) algorithm but still having a good variance reduction in the controlled responses of the simulation. The SA has performed successfully as a general heuristic algorithm for the solution of large, complex combinatorial optimization problems. For more information of the SA algorithm, see kirkpatrick et al. (1983).

Table 4. Internal variance estimators when ACVs are applied.

n	$\widehat{Var}[\bar{Y}]$	$\widehat{Var}[\bar{Y}_W(ALL)]$	$\widehat{Var}[\bar{Y}_W(ACV)]$	$VR_{ALL}(\%)$	$VR_{ACV}(\%)$	$\left[\frac{VR_{ACV}}{VR_{ALL}}\right]$
11	0.9429	0.5698	0.2398	40	75	1.88
13	0.8068	0.3435	0.2317	57	71	1.25
15	0.7020	0.2935	0.2256	58	68	1.17
18	0.5860	0.2265	0.2047	61	65	1.07
20	0.5290	0.1954	0.1722	63	67	1.06
23	0.4536	0.1628	0.1502	64	67	1.05
25	0.4165	0.1540	0.1452	63	65	1.03
28	0.3752	0.1363	0.1309	64	65	1.02

4.2 Experimental Results Using the ACV Method

Since it has been shown in the previous section that using **W** is more efficient than using **V**, we applied **W** only for the experiments in this section. Table 4 shows the difference in internal variance estimators (IVEs) of the simulation responses among three cases. The number of simulation runs is in the first column. Variance estimators without using any control variates and using all 8 control variates are shown in the second and the third columns, respectively. Finally, the variance estimators computed by using the ACV method are in the fourth column. The percentages of variance reduction are given in fifth and sixth columns. All variance estimators in Table 4 are obtained from a macroexperiment consisting of 100 microexperiments as discussed in previous section. The constant value c of the ACV procedure in Figure 3 is a key parameter which is supposed to be prespecified. In this experiment, we use the value of c such that $c \times \hat{R}^2_{\max}$=0.9 because we assume that 90 percents of $\hat{R}^2$ is the most efficient cutoff-point to determine the number of CVs. However, we found that the results of ACV can be slightly improved by adjusting the c value depending upon $\hat{R}^2_{\max}$ and the number of simulation runs. A precise mathematical relationship among these parameters should be developed in the future. Note that as n increases, the efficiency of ACV decreases. That is because the loss factor is vanishing as n increases. In this experiment, if n 〉 30 the advantage of using the ACV method seems to be negligible compared to the case of using all control variates. See the decreasing trend in the last column (Table 4), which shows the ratio of the percentage variance reduction for the ACV method divided by the percentage of variance reduction for the case in which all CVs were used. Therefore, the ACV method is useful particularly when a high precision in the responses of the simulation is not required.

Table 5 : External variance estimators when ACVs are applied.

n	$Var[\bar{Y}]$	$Var[\bar{Y}_W(allCVs)]$	$Var[\bar{Y}_W(ACV)]$
11	0.8817	3.1320	1.8266
13	0.7120	0.7851	0.6590
15	0.7235	0.5217	0.4367
18	0.5295	0.3183	0.3085
20	0.5392	0.3093	0.3032
23	0.4830	0.3028	0.3306
25	0.4653	0.1919	0.1965
28	0.4514	0.1652	0.1872

In Table 5, we also compared the external variance estimators (EVEs) with the same input data used in Table 4. However, there is a problem in the actual application of ACV to the SA algorithm. That is, in order to obtain EVE with the SA algorithm, we need a relatively long simulation run to provide the batch means and variances to compute EVEs at each temperature. This long simulation run at each temperature will prevent the SA algorithm from speeding up. Moreover, at each temperature, it is not easy to decide how many simulation runs should be performed to meet a prespecified precision. If we use IVE, however, it is easy to decide the number of simulation runs required to meet the precision.

Another problem for using EVE is its instability. In other words, EVE does not steadily decrease as n increases. For examples, see the EVEs for $18 \leq n \leq 20$ in the second column and the EVEs for $20 \leq n \leq 23$ in the fourth column of Table 5. When $n = 11$, Table 5 shows that controlled EVEs for ACV (1.8266) and for the case of all CVs (3.1320) are much greater than for the case of no controls (0.8817). This never happens in computing IVEs because of step [5] in Figure 3. Therefore, in the method of ACV proposed in this paper, we recommend to use IVEs in all experiments with an actual application of ACV to the SA algorithm in the optimization of a queueing network simulation.

5. Conclusion

We developed a method of adaptive control variates (ACVs) to reduce the number of simulation runs required at each temperature of the SA algorithm. In the ACV procedure, an optimal subset of CVs is selected automatically. Moreover, since the SA algorithm requires iterated simulation runs in its application to most stochastic combinatorial optimization problems (SCOPs), variance reduction is critical to the true optimal solution in a given queueing network system. Therefore, the application of SA to the optimization of queueing network simulations is an excellent example to show how a variance reduction technique (VRT) can drastically improve the performance of SA in solving stochastic optimization problems.

References

[1] Anderson, T.W.(1984). An Introduction to Multivariate Statistical Analysis, Ind edition . John Wiley & Sons, New York.

[2] Añonuevo, R. and Nelson, B.L. (1988), "Automated estimation and variance reduction via control variates for infinite-horizon simulations", Comput. Operations Research 15, 447-456.

[3] Bauer, K.W. (1987), "Control variate selection for multiresponse simulation", Ph.D. Dissertation, School of Industrial Engineering, Purdue University, West Lafayette, IN 47907.

[4] Bauer, K.W. and Wilson, J.R. (1992), "Control variate selection criteria", Naval Research Logistics 39, 307-321.

[5] Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi. (1983). "Optmization by simulated annealing," Science, Vol.220, pp.671-680.

[6] Lavenberg, S.S., Moeller, T.L., and Welch, P.D. (1982), "Statistical results on control variables with appication to queueing network simulation", Operations Research 30, 182-202.

[7] Lee, J.Y. (1995), "Faster simulated annealing techniques for stochastic optimization problems, with application to queueing network simulation", Ph.D. Dissertation, Statistics and Operations Research in North Carolina State University, Raleigh, NC 27695.

[8] Nelson, B.L. (1987), "A perspective on variance reduction in dynamic simulation experiments", Communications in Statistics B 16, 385-426.

[9] Neter, J., W. Wasserman, and M.H. Kutner. (1990), "Applied linear statistical models, 3rd edition", Irwin, Homewood, IL 60430.

[10] Nozari, A., Arnold, S.F., and Pegden, C.D. (1984), "Control variates for multipopulation simulation experiments", IIE Transactions 16, 159-169.

[11] Porta Nova, A.M. and Wilson, J.R. (1993), "Selection control variates to estimate multiresponse simulation metamodels", European J. of Operations Research 71, 80-94.

[12] Rubinstein, R.Y. and Marcus, R. (1985), "Efficiency of multivariate control variates in Monte Carlo simulation", Operations Research 33, 661-677.

[13] Wilson, J.R. (1984), "'Variance reduction techniques for digital simulation" American J. of Mathematical and Management Science, Vol.4, 277-312.

[14] Wilson, J.R., and A.A.B. Pritsker. (1984), "Variance

reduction in qeueing simulation using generalized concomitant variables", J. of Statistical Computation and Simulation, Vol.19, pp.129-153.

● 저자소개 ●

이재영

1980. 3 육군사관학교 졸업

1988. 9 미국 해군대학원 OR석사 과정 졸업

1995.12 미국 North Carolina 주립대학 통계 및 OR박사 과정 졸업(복수전공)

현재 육군 교육사령부 BCTP단 DB계획장교