Proceedings of the 1986 Winter Simulation Conference
J. Wilson, J. Henriksen, S. Roberts (eds.)

USING CONTROL VARIATES TO ESTIMATE MULTIRESPONSE SIMULATION METAMODELS

Acácio M. de O. Porta Nova
Departamento de Matemática
Instituto Superior Técnico
Avenida Rovisco Pais
1096 Lisboa CODEX, Portugal

James R. Wilson
School of Industrial Engineering
Purdue University
West Lafayette, IN 47907, U.S.A.

ABSTRACT

In this paper we extend the method of control variates to the estimation of a multiresponse simulation metamodel—that is, a multivariate linear model of selected simulation responses expressed in terms of relevant decision variables for the target system. For the case in which the responses and the controls have a joint normal distribution, we present control variates point and confidence region estimators of the coefficients of a multiresponse metamodel, and we describe a procedure for testing a general linear hypothesis about the metamodel. To quantify the maximum efficiency that is achievable with a given set of controls, we introduce a generalized minimum variance ratio. To measure the degradation in efficiency that occurs when the optimal control coefficients are estimated, we formulate a generalized loss factor. An example illustrates the application of these results.

1. INTRODUCTION

Since the advent of the digital computer, simulation has made possible the study of very large scale systems. However, direct simulation of a realistic model of such a system can still be prohibitively expensive, either by requiring too many simulation runs or by requiring excessive run lengths. A variety of variance reduction techniques (VRTs) have been proposed to improve the efficiency of simulation experiments; see Kleijnen (1974) or Wilson (1984) for a comprehensive survey of these techniques.

In the simplest form of a simulation experiment, we assume that each of n simulation runs produces an independent and identically distributed response Y whose mean $\mu_Y = E(Y)$ is to be estimated. The *direct simulation estimator* of μ_Y is the sample mean response $\overline{Y}$ taken over all n replications. This estimator is unbiased so that $E(\overline{Y}) = \mu_Y$, and it has variance $Var(\overline{Y}) = Var(Y)/n$. The *variance reduction problem* consists of finding another unbiased estimator of μ_Y with a smaller variance. The *method of control variates* is a VRT that exploits inherent linear correlation between the target response Y and a concomitant output variate C (a "control") that is also observed on each run and that has a *known* mean μ_C. On the i^{th} run of the simulation model, we compute the "controlled" response $Y_i(\phi) = Y_i - \phi(C_i - \mu_C)$, thereby attempting to compensate for the unknown estimation error $Y_i - \mu_Y$ by subtracting from the original response Y_i a linear transformation of the corresponding known deviation $C_i - \mu_C$. The "controlled" estimator of μ_Y is simply the sample mean $\overline{Y}(\phi) = \overline{Y} - \phi(\overline{C} - \mu_C)$ of these "controlled" responses.

For any fixed value of the control coefficient ϕ, the controlled response $Y(\phi) = Y - \phi(C - \mu_C)$ is unbiased with variance

$$Var[Y(\phi)] = Var(Y) + \phi^2 Var(C) - 2\phi Cov(Y, C). \quad (1)$$

The method of control variates yields a variance reduction relative to direct simulation if ϕ lies between 0 and $2Cov(Y, C)/Var(C)$. If the covariance structure of the system under study is known, then calculus can be used to determine the value of ϕ that minimizes (1). The optimal control coefficient is $\delta = Cov(Y, C)/Var(C)$; and the resulting minimum variance achievable with the control C is $Var[Y(\delta)] = Var(Y)(1-\rho_{YC}^2)$, where ρ_{YC} is the *correlation coefficient* between Y and C. In general, δ must be estimated, and its least-squares estimator, $\hat{\delta}$, is the sample equivalent of the optimal control coefficient, δ, defined above. When Y and C are jointly normal and the least-squares estimator $\hat{\delta}$ is used for δ, the variance ratio

$$VR(\hat{\delta}) \equiv \frac{Var[\overline{Y}(\hat{\delta})]}{Var(\overline{Y})} = \left(\frac{n-2}{n-3}\right)(1-\rho_{YC}^2)$$

is actally achieved with the control C. The factor $(n-2)/(n-3)$ measures the loss of efficiency that results from estimating δ; see Porta Nova (1985).

In contrast to other variance reduction techniques (for instance, those that are based on the principle of importance sampling), the method of control variates does not require manipulation or distortion of the random number

input process; instead, this latter method is based on observation of the natural dynamic behavior of the simulated system. In a wide variety of applications, the control variates technique can yield large variance reductions with little additional computing overhead. These considerations suggest that control variates have great potential for actual application in simulation studies of real systems.

2. SETUP FOR METAMODEL ESTIMATION

2.1. Notation and Assumptions

Although we must introduce some new notation to accomodate our extensions of the control variates technique, we have attempted to incorporate much of the symbolism that is common to recent papers in the field—in particular Lavenberg, Moeller, and Welch (1982); Nozari, Arnold, and Pegden (1984); Rubinstein and Marcus (1985); and Venkatraman and Wilson (1986). For an $s \times t$ matrix $\mathbf{A} = ||A_{ij}||$, we let $\mathbf{A}_{i\cdot}$ denote the i^{th} row vector of $\mathbf{A}$ $(1 \le i \le s)$, and we let $\mathbf{A}_{\cdot j}$ denote the j^{th} column vector of $\mathbf{A}$ $(1 \le j \le t)$. Thus we have $\mathbf{A} = [\mathbf{A}_{\cdot 1} \cdots \mathbf{A}_{\cdot t}] = [\mathbf{A}_{1\cdot}' \cdots \mathbf{A}_{s\cdot}']'$. Also, vec(A) or vec $\mathbf{A}$ denotes the st-dimensional column vector obtained by stacking the columns of $\mathbf{A}$ respectively under one another to form a single column: $\text{vec}(\mathbf{A}) = \text{vec}\,\mathbf{A} \equiv [\mathbf{A}_{\cdot 1}' \cdots \mathbf{A}_{\cdot t}']'$. If $\mathbf{B}$ is a $u \times v$ matrix, then the *right direct product* of $\mathbf{A}$ and $\mathbf{B}$ is the $su \times tv$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} A_{11}\mathbf{B} & \cdots & A_{1t}\mathbf{B} \\ \cdot & \cdots & \cdot \\ A_{s1}\mathbf{B} & \cdots & A_{st}\mathbf{B} \end{bmatrix}.$$

See Searle (1982) for an elaboration of these definitions of vec(A) and $\mathbf{A} \otimes \mathbf{B}$.

We assume that n independent simulation runs have been performed in order to produce independent observations of each of the design points of a chosen experimental layout. For the i^{th} simulation run (i=1, 2, ..., n), we define the following quantities: (a) a $1 \times m$ deterministic vector $\mathbf{X}_{i\cdot} = [X_{i1}, ..., X_{im}]$ of decision variables or input parameters for the simulation model, (b) a $1 \times p$ random vector $\mathbf{Y}_{i\cdot} = [Y_{i1}, ..., Y_{ip}]$ of simulation responses, and (c) a $1 \times q$ random vector $\mathbf{C}_{i\cdot} = [C_{i1}, ..., C_{iq}]$ of concomitant control variables with a known mean. Thus the $n \times m$ matrix $\mathbf{X}$ specifies the entire experimental layout, the $n \times p$ matrix $\mathbf{Y}$ represents all the system responses of interest, and the $n \times q$ matrix $\mathbf{C}$ contains all the information about the relevant control variables observed during the experiment.

We make the following assumptions about the joint distribution of the responses and the controls. The control vector $\mathbf{C}_{i\cdot}$ observed on each run is assumed to have mean $\mathbf{0}_{1\times q}$ and the same probability density function at all design points. This last property ensures that the covariance matrix $\text{Cov}(\mathbf{C}_{i\cdot}) = \underset{\sim}{\Sigma}_C$ is positive definite (p.d.) and constant across all design points; see Porta Nova (1985). The response $\mathbf{Y}_{i\cdot}$ is assumed to be given by the multivariate linear model $\mathbf{Y}_{i\cdot} = \mathbf{X}_{i\cdot}\underset{\sim}{\Theta} + \underset{\sim}{\epsilon}_{i\cdot}$, $i = 1, ..., n$, where $\underset{\sim}{\Theta}$ is the $m \times p$ matrix of unknown metamodel coefficients and $\underset{\sim}{\epsilon}_{i\cdot}$ is a $1 \times p$ vector of errors with $E(\underset{\sim}{\epsilon}_{i\cdot}) = \mathbf{0}_{1\times p}$, $i = 1, ..., n$. The covariance matrix of $\mathbf{Y}_{i\cdot}$ (and of $\underset{\sim}{\epsilon}_{i\cdot}$) is assumed to be p.d. and constant across all design points, $\text{Cov}(\mathbf{Y}_{i\cdot}) = \underset{\sim}{\Sigma}_Y$, $i = 1, ..., n$. The covariance matrix between $\mathbf{Y}_{i\cdot}$ and $\mathbf{C}_{i\cdot}$ is assumed to be constant across all design points, $\text{Cov}(\mathbf{Y}_{i\cdot}, \mathbf{C}_{i\cdot}) = \underset{\sim}{\Sigma}_{YC} = \underset{\sim}{\Sigma}_{CY}'$, $i = 1, ..., n$. Finally we assume that the error $\underset{\sim}{\epsilon}_{i\cdot} = \mathbf{Y}_{i\cdot} - \mathbf{X}_{i\cdot}\underset{\sim}{\Theta}$ has a linear regression on $\mathbf{C}_{i\cdot}$ with an unknown $q \times p$ matrix $\underset{\sim}{\Delta}$ of regression (control) coefficients and with a $1 \times p$ matrix of residuals $\mathbf{R}_{i\cdot}$ so that the equation

$$\mathbf{Y}_{i\cdot} = \mathbf{X}_{i\cdot}\underset{\sim}{\Theta} + \mathbf{C}_{i\cdot}\underset{\sim}{\Delta} + \mathbf{R}_{i\cdot}, \quad i = 1, ..., n, \tag{2}$$

is a valid metamodel for the overall simulation experiment.

2.2. Objectives

In this paper we present two sets of results:

1. Assuming the validity of the model (2), we present the following: (a) a generalization of the minimum variance ratios formulated by Lavenberg, Moeller, and Welch (1982) and Rubinstein and Marcus (1985); (b) the least squares estimator $\hat{\underset{\sim}{\Delta}}$ for the optimal control coefficient matrix $\underset{\sim}{\Delta}$; and (c) the corresponding least squares estimator $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$ for the metamodel coefficient matrix $\underset{\sim}{\Theta}$ in (2).

2. With the additional assumption that the response vector $\mathbf{Y}_{i\cdot}$ and the control vector $\mathbf{C}_{i\cdot}$ have a joint (p+q)-dimensional normal distribution, we present the following: (a) an exact $100(1-\alpha)\%$ confidence ellipsoid for $\text{vec}\underset{\sim}{\Theta}$ centered at $\text{vec}\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$; (b) generalizations of the loss factors derived by Lavenberg, Moeller, and Welch (1982) and Venkatraman and Wilson (1986); and (c) a

procedure for testing the general linear hypothesis H_0: $\mathbf{G}\,\mathrm{vec}\underset{\sim}{\Theta} = \mathbf{b}$, where $\mathbf{G}$ is an $s \times mp$ matrix of full row rank and $\mathbf{b}$ is a given $s \times 1$ vector.

To evaluate the performance of the estimation procedures and efficiency measures derived under the normality assumption, we conducted an extensive simulation study of the basic experiment described in the next section. Some of the numerical results of this study are presented in Section 6.

3. DESCRIPTION OF AN APPLICATION

We will apply the methodology introduced in this paper to an example that was originally presented by Lavenberg, Moeller, and Welch (1982). The target system consists of a closed queueing network model representing an interactive multiprogrammed computer system with multiple customer classes and a subnetwork capacity constraint. Lavenberg, Moeller, and Welch analyzed variations of this basic system in the context of a univariate response ($p=1$) with multiple controls ($q \geq 1$), where the objective was to estimate the overall mean response ($m=1$). In our context, we seek to estimate a linear model for a certain performance measure computed within each customer class ($p \geq 1$), where the metamodel for each class is expressed in terms of several relevant system design parameters ($m \geq 1$); moreover, we want to make efficient use of all available controls ($q \geq 1$). Figure 1 shows the particular queueing system to be analyzed.

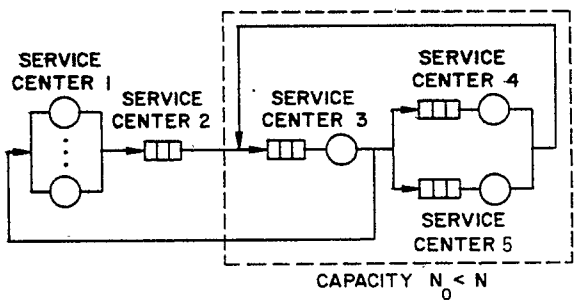

Figure 1: Computer System Model with Subsystem Capacity Constraint

There are two classes of customers using the computer with N(k) customers in class k, $k=1, 2$. Service center 1 has N(1)+N(2) servers representing the terminals, so no queueing occurs at center 1. The service time at center 1 represents a customer's "think" time at the terminal. Service centers 3, 4, and 5 are single-server, first-in, first-out (FIFO) queues that respectively model a central processor (center 3) and its associated secondary storage devices (centers 4 and 5). Centers 3, 4, and 5 constitute a subnetwork with capacity N_0, meaning that no more than N_0 jobs can be multiprogrammed by the computer system at the same time. A class k job that finishes service at center 3 is routed to service center j with probability $p_j(k)$, where $k=1, 2$ and $j=1, 4, 5$. Service times at center 3 represent processing times for jobs. Service times at centers 4 and 5 represent times to access and transfer information between the main memory and the secondary storage devices. The queue at service center 2 holds those jobs waiting to be activated when there are already N_0 active jobs in the subnetwork. Center 2 represents a communications processor that holds jobs until enough main memory is available. At center 2, class 1 jobs have higher priority and all service times are identically zero. Table 1 summarizes all of the parameters describing this system.

Table 1: Parameters of the Computer System Model

Parameters	Class k = 1	Class k = 2
Number of Users N(k)	15	10
Routing probability $p_j(k)$ from center 3 to center j for class k:		
$p_1(k)$	0.2	0.2
$p_4(k)$	0.72	0.4
$p_5(k)$	0.08	0.4
Mean service time $\mu_{j,k}$ at center j for class k:		
$\mu_{1,k}$	100.0	100.0
$\mu_{3,k}$	1.0	1.0
$\mu_{4,k}$	0.694	0.694
$\mu_{5,k}$	6.25	6.25
Subsystem capacity constraint: $N_0 = 5$		

The purpose of our simulation experiment is to fit a separate linear model to the mean response time for each customer class in terms of three decision variables: the number of customers in each class and the capacity of the central processor subnetwork. A *response time* is the delay between the departure of a job from service center 1 and the job's subsequent return to center 1. We are interested in exploring the region of the decision space around the point $N(1) = 15$, $N(2) = 10$ and $N_0 = 5$. The proposed model is a second-degree polynomial in the decision variables. To estimate this model, we take 3 replicates of each

point in a simple 3^3 factorial design. The values of each decision variable are coded as 0 (low level), 1 (intermediate level) or 2 (high level). Table 2 summarizes the experimental design used to estimate the bivariate model of response times.

Table 2: Experimental Layout for the Computer System Simulation

Decision Variable	Original Values			Coded Values†			Symbol in Bivariate Model
	L	I	H	L	I	H	
N(1)	12	15.	18	0	1	2	X_1
N(2)	8	10	12	0	1	2	X_2
N_0	3	5	7	0	1	2	X_3

†Each treatment combination is replicated three times.

On the i^{th} simulation run (i=1, ..., 81), the proposed model for the expected value of the vector $\mathbf{Y}_{i\cdot} = [Y_{i1}, Y_{i2}]$ of sample average response times for customer classes 1 and 2 is given by

$$\mathbf{Y}_{i\cdot} = \mathbf{X}_{i\cdot}\underset{\sim}{\Theta}, \tag{3}$$

where $\mathbf{X}_{i\cdot} =$

$$\left[1 \;\; X_{i1} \;\; X_{i2} \;\; X_{i3} \;\; X_{i1}^2 \;\; X_{i2}^2 \;\; X_{i3}^2 \;\; X_{i1}X_{i2} \;\; X_{i1}X_{i3} \;\; X_{i2}X_{i3}\right]$$

and $\underset{\sim}{\Theta} = ||\Theta_{ij}||$ is the 10×2 matrix of metamodel coefficients. Thus in this example we have p=2, m=10, and n=81.

To obtain a more precise estimator of the postulated bivariate metamodel (3), we apply a set of standardized control variates proposed by Venkatraman (1983). For service center j and customer class k,

$$C_{j,k}^*(t) = t^{-1/2} \sum_{l=1}^{g(j,\,k;\,t)} (U_{j,k,l} - \mu_{j,k})/\sigma_{j,k},$$

where: k=1, 2 is the index of the customer class; j=3, 4, 5 is the index of the service center; g(j, k; t) is the number of service times started at station j for customer class k during the simulated time interval (0, t]; and $U_{j,k,l}$ is the l^{th} service time sampled at station j for customer class k, where $E(U_{j,k,l}) = \mu_{j,k}$ and $Var(U_{j,k,l}) = \sigma_{j,k}^2$. Thus if T is a fixed simulation ending time, then the vector of controls observed at the i^{th} design point is given by

$$\mathbf{C}_{i\cdot} = \left[C_{3,1}^*(T) \;\; C_{4,1}^*(T) \;\; C_{5,1}^*(T) \;\; C_{3,2}^*(T) \;\; C_{4,2}^*(T) \;\; C_{5,2}^*(T)\right]$$

where the $\{C_{j,k}^*(T)\}$ are of course accumulated on the i^{th} run. We do not use $C_{1,1}^*(T)$ as a control because the response time for a job does not include the "think" time at the terminal. In contrast, the service times at the other centers enter into the evaluation of the response times; and it is reasonable to assume that there will be some correlation between the response times and the service times sampled at centers 3, 4 and 5. Thus q=6 in this example.

4. SUMMARY OF NONPARAMETRIC RESULTS

In this section we present the results on controlled estimation of multiresponse metamodels that do not depend on the assumption of joint normality for the responses and the controls. In Section 5 we summarize the main results on estimation and hypothesis testing for multinormal metamodels. Proofs of all of these results can be found in Porta Nova (1985) and in Porta Nova and Wilson (1986).

4.1. Minimum Variance Ratio

In terms of the aggregate experimental data matrices $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{C}$, and $\underset{\sim}{\varepsilon}$, the linear model that we are estimating is compactly expressed as $\mathbf{Y} = \mathbf{X}\underset{\sim}{\Theta} + \underset{\sim}{\varepsilon}$. The design matrix $\mathbf{X}$ is assumed to have full column rank m so that the ordinary least squares estimator of $\underset{\sim}{\Theta}$ is $\hat{\underset{\sim}{\Theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. In terms of the vec operator, the covariance matrix of $\hat{\underset{\sim}{\Theta}}$ can be conveniently expressed as

$$Cov(vec\hat{\underset{\sim}{\Theta}}) = \underset{\sim}{\Sigma}_Y \otimes (\mathbf{X}'\mathbf{X})^{-1};$$

and this implies that the generalized variance of $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$ is

$$|\, Cov(vec\hat{\underset{\sim}{\Theta}})\,| \;=\; |\,\underset{\sim}{\Sigma}_Y\,|^{\,m}\,|\,\mathbf{X}'\mathbf{X}\,|^{\,-p}.$$

To obtain a more efficient estimator for $\underset{\sim}{\Theta}$, we attempt to predict the unobservable error $\underset{\sim}{\varepsilon}$ in the response $\mathbf{Y}$ as a linear transformation $\mathbf{C}\underset{\sim}{\Phi}$ of the control $\mathbf{C}$; thus the component of $\underset{\sim}{\varepsilon}$ that is linearly associated with $\mathbf{C}$ can be removed from $\mathbf{Y}$ before computing the least-squares estimator of $\underset{\sim}{\Theta}$. For any fixed q×p matrix $\underset{\sim}{\Phi}$ of control coefficients, the controlled estimator of $\underset{\sim}{\Theta}$ is

$$\hat{\underset{\sim}{\Theta}}(\underset{\sim}{\Phi}) \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{C}\underset{\sim}{\Phi}),$$

which is seen to be unbiased with generalized variance

$$|\,\mathrm{Cov}[\mathrm{vec}\hat{\underset{\sim}{\Theta}}(\underset{\sim}{\Phi})]\,| =$$

$$|\,\underset{\sim}{\Sigma}_Y + \underset{\sim}{\Phi}'\underset{\sim}{\Sigma}_C\underset{\sim}{\Phi} - \underset{\sim}{\Phi}'\underset{\sim}{\Sigma}_{CY} - \underset{\sim}{\Sigma}_{YC}\underset{\sim}{\Phi}\,|^{m}\,|\,\mathbf{X}'\mathbf{X}\,|^{-p}.$$

For the optimal control coefficient matrix $\underset{\sim}{\Delta} = \underset{\sim}{\Sigma}_C^{-1}\underset{\sim}{\Sigma}_{CY}$, the minimum variance ratio is

$$\mathrm{VR}(\underset{\sim}{\Delta}) \equiv \frac{|\,\mathrm{Cov}[\mathrm{vec}\underset{\sim}{\Theta}(\underset{\sim}{\Delta})]\,|}{|\,\mathrm{Cov}(\mathrm{vec}\underset{\sim}{\Theta})\,|} = \left[\prod_{j=1}^{\nu}(1-\rho_j^2)\right]^m, \quad (4)$$

where ν denotes the rank of $\underset{\sim}{\Sigma}_{YC}$ and $\{\rho_j : j = 1, ..., \nu\}$ are the canonical correlations between $\mathbf{Y}$ and $\mathbf{C}$. This is a natural generalization of the corresponding efficiency measures defined by Lavenberg, Moeller, and Welch (1982) and Rubinstein and Marcus (1985). If $\nu = 0$, then we take $[\prod(1-\rho_j^2)] \equiv 1$ in equation (4). For a univariate response (p=1) with $\nu = 1$, we see that ρ_1 is the coefficient of multiple correlation between $\mathbf{Y}$ and $\mathbf{C}$. The maximum percentage reduction in generalized variance that can be obtained with the control $\mathbf{C}$ is 100[1–VR($\underset{\sim}{\Delta}$)]. Of course, we do not know the optimal control coefficient matrix $\underset{\sim}{\Delta}$ in general, and a loss factor is required to quantify the subsequent percentage increase in generalized variance that occurs when the optimal control coefficients must be estimated.

4.2. Least-Squares Metamodel Estimator

The procedure used to obtain an estimator for the optimal control coefficient matrix $\underset{\sim}{\Delta}$ and to obtain the controlled estimator for the metamodel coefficient matrix $\underset{\sim}{\Theta}$ is a generalization of the least squares (LS) method based on matrix derivatives. The LS estimator for $\underset{\sim}{\Delta}$ is

$$\hat{\underset{\sim}{\Delta}} = (\mathbf{C}'\mathbf{P}\mathbf{C})^{-1}\mathbf{C}'\mathbf{P}\mathbf{Y}, \text{ where } \mathbf{P} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'; \quad (5)$$

and the corresponding estimator for $\underset{\sim}{\Theta}$ is

$$\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{C}\hat{\underset{\sim}{\Delta}}). \quad (6)$$

The controlled estimator (6) does not appear to resemble the one described in our basic framework of Section 1, namely $\overline{Y}(\hat{\delta}) = \overline{Y} - \hat{\delta}(\overline{C} - \mu_C)$. However, if we take $\mathbf{Y} = [Y_1, ..., Y_n]'$, $\mathbf{X} = \mathbf{1}_{n\times 1}$ (an n×1 vector of ones), and $\mathbf{C} = [C_1-\mu_C, ..., C_n-\mu_C]'$, then we see that $\hat{\underset{\sim}{\Delta}} = \hat{\delta}$, the sample covariance between the $\{Y_i\}$ and the $\{C_i\}$ divided by the sample variance of the $\{C_i\}$; and $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}}) = \overline{Y}(\hat{\delta})$. So, we see that (6) is the direct generalization of the "controlled" estimator concept introduced by Lavenberg, Moeller, and Welch (1982).

5. RESULTS FOR NORMAL METAMODELS

The n×(p+q) matrix $\mathbf{Z}$ is said to be normally distributed with the mean matrix $\underset{\sim}{\mu}_Z = ||\mu_{ij}||$, the covariance matrix between rows $\underset{\sim}{\Xi} = ||\Xi_{ij}||$, and the covariance matrix between columns $\underset{\sim}{\Sigma}_Z = ||\Sigma_{ij}||$ if: (a) each element Z_{ij} is normally distributed with mean $E(Z_{ij}) = \mu_{ij}$; (b) the covariance between $\mathbf{Z}_{i\cdot}$ and $\mathbf{Z}_{j\cdot}$ is $\mathrm{Cov}(\mathbf{Z}_{i\cdot}, \mathbf{Z}_{j\cdot}) = \Xi_{ij}\underset{\sim}{\Sigma}_Z$ for i, j = 1, ..., n; and (c) the covariance between $\mathbf{Z}_{\cdot k}$ and $\mathbf{Z}_{\cdot l}$ is $\mathrm{Cov}(\mathbf{Z}_{\cdot k}, \mathbf{Z}_{\cdot l}) = \Sigma_{kl}\underset{\sim}{\Xi}$, for k, l = 1, ..., p+q. We let

$$\mathbf{Z} \sim \mathrm{N}_{n,p+q}(\underset{\sim}{\mu}_Z, \underset{\sim}{\Xi}, \underset{\sim}{\Sigma}_Z)$$

denote this matrix normal distribution. (See Section 17.2 of Arnold (1981) for an elaboration of this definition.)

For a simulation experiment in which the response matrix $\mathbf{Y}$ and the control matrix $\mathbf{C}$ jointly possess a normal distribution, we take $\mathbf{Z} \equiv (\mathbf{Y}, \mathbf{C})$ so that $E(\mathbf{Z}) = \underset{\sim}{\mu}_Z = (\underset{\sim}{\mu}_Y, \mathbf{0}_{n\times q})$, with $\underset{\sim}{\mu}_Y = \mathbf{X}\underset{\sim}{\Theta}$, and

$$\underset{\sim}{\Sigma}_Z = \begin{bmatrix} \underset{\sim}{\Sigma}_Y & \underset{\sim}{\Sigma}_{YC} \\ \underset{\sim}{\Sigma}_{CY} & \underset{\sim}{\Sigma}_C \end{bmatrix}.$$

We assume that $\underset{\sim}{\Xi}$ and $\underset{\sim}{\Sigma}_Z$ are positive definite. We also assume that the rows of $\mathbf{Z}$ are mutually independent since they correspond to independent simulation runs; thus we take $\underset{\sim}{\Xi} = \mathbf{I}_n$ in the following development.

Although we have derived the corresponding results for the case when $\underset{\sim}{\Sigma}_Z$ is known (see Porta Nova (1985)), in this paper we consider only the more realistic situation in which $\underset{\sim}{\Sigma}_Z$ must be estimated. If $\mathbf{Z} \sim \mathrm{N}_{n,p+q}(\underset{\sim}{\mu}_Z, \mathbf{I}_n, \underset{\sim}{\Sigma}_Z)$ with both $\underset{\sim}{\Theta}$ and $\underset{\sim}{\Sigma}_Z$ unknown, then we exploit the fact that the conditional distribution of $\mathbf{Y}$ given $\mathbf{C}$ is matrix normal,

$$\mathbf{Y} \mid \mathbf{C} \sim \mathrm{N}_{n,p}(\underset{\sim}{\mu}_{Y\cdot C}, \mathbf{I}_n, \underset{\sim}{\Sigma}_{Y\cdot C}), \quad \text{with}$$

$$\underset{\sim}{\mu}_{Y\cdot C} = \underset{\sim}{\mu}_Y + \mathbf{C}\underset{\sim}{\Sigma}_C^{-1}\underset{\sim}{\Sigma}_{CY}, \quad \underset{\sim}{\Sigma}_{Y\cdot C} = \underset{\sim}{\Sigma}_Y - \underset{\sim}{\Sigma}_{YC}\underset{\sim}{\Sigma}_C^{-1}\underset{\sim}{\Sigma}_{CY}.$$

Thus by conditioning on $\mathbf{C}$, we see that the assumed linear model (2) of Section 2.2 is in fact the correct model for the matrix normal response $\mathbf{Y}$, where: (a) the control coefficient matrix is $\underset{\sim}{\Delta} = \underset{\sim}{\Sigma}_C^{-1}\underset{\sim}{\Sigma}_{CY}$; and (b) the residual matrix $\mathbf{R}$ is independent of $\mathbf{C}$ with

$$\mathbf{R} \sim \mathrm{N}_{n,p}(\mathbf{0}_{n\times p}, \mathbf{I}_n, \underset{\sim}{\Sigma}_{Y\cdot C}).$$

5.1. Distribution of the Metamodel Estimator

Given the control matrix $\mathbf{C}$, the least-squares estima-

tor $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$ is matrix normally distributed with

$$\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}}) \mid \mathbf{C} \sim N_{m,p}(\mathbf{B}\underset{\sim}{\mu}_{Y\cdot C}, \mathbf{BB}', \underset{\sim}{\Sigma}_{Y\cdot C}), \quad \text{where}$$

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[\mathbf{I}_n - \mathbf{C}(\mathbf{C}'\mathbf{P}\mathbf{C})^{-1}\mathbf{C}'\mathbf{P}\right].$$

We have $\mathbf{B}\underset{\sim}{\mu}_{Y\cdot C} = \underset{\sim}{\Theta}$, showing the conditional unbiasedness of $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$ given $\mathbf{C}$ (and hence also the unconditional unbiasedness of $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$). An unbiased estimator for $\underset{\sim}{\Sigma}_{Y\cdot C}$ is

$$\hat{\underset{\sim}{\Sigma}}_{Y\cdot C} = \mathbf{Y}'(\mathbf{P}-\mathbf{Q})\mathbf{Y}/(n-m-q), \text{ and } \mathbf{Q} = \mathbf{PC}(\mathbf{C}'\mathbf{PC})^{-1}\mathbf{C}'\mathbf{P}.$$

Conditioned on $\mathbf{C}$, $\mathbf{Y}'(\mathbf{P}-\mathbf{Q})\mathbf{Y}$ has a p-dimensional (central) Wishart distribution with n-m-q degrees of freedom and covariance matrix $\underset{\sim}{\Sigma}_{Y\cdot C}$. Moreover, given $\mathbf{C}$, $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$ and $\hat{\underset{\sim}{\Sigma}}_{Y\cdot C}$ are conditionally independent random matrices.

5.2. Variance Ratio and Loss Factor

Stacking the columns of the estimator $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$ into the mp×1 column vector $\text{vec}\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$, we have:

$$\text{Var}[\text{vec}\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})] = (\underset{\sim}{\Sigma}_Y - \underset{\sim}{\Sigma}_{YC}\underset{\sim}{\Sigma}_C^{-1}\underset{\sim}{\Sigma}_{CY}) \otimes \frac{n-m-1}{n-m-q-1}(\mathbf{X}'\mathbf{X})^{-1}.$$

When $\underset{\sim}{\Delta}$ must be estimated, the actual variance ratio is

$$\text{VR}(\hat{\underset{\sim}{\Delta}}) = \frac{|\text{Var}[\text{vec}\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})]|}{|\text{Var}(\text{vec}\hat{\underset{\sim}{\Theta}})|} = \left(\frac{n-m-1}{n-m-q-1}\right)^p \cdot \left[\prod_{j=1}^{\nu}(1-\rho_j^2)\right]^m. \quad (7)$$

Thus, when we must estimate the optimal control coefficient matrix $\underset{\sim}{\Delta}$, the actual variance ratio $\text{VR}(\hat{\underset{\sim}{\Delta}})$ results from taking the product of the minimum variance ratio (4) and the loss factor

$$\lambda = \left(\frac{n-m-1}{n-m-q-1}\right)^p. \quad (8)$$

We see that if the number of replications is very large relative to the number of parameters to be estimated ($n \gg m+q$), then the loss factor becomes unimportant, $\lambda \approx 1$. However, if n is small relative to m+q, then a net variance increase (rather than a net variance reduction) can occur.

5.3. A Confidence Region for the Metamodel

Let $\mathbf{V} \equiv \hat{\underset{\sim}{\Sigma}}_{Y\cdot C} \otimes \mathbf{BB}'$. From the results cited in Section 5.1, it follows that

$$\text{vec}[\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}}) - \underset{\sim}{\Theta}]' \mathbf{V}^{-1} \text{vec}[\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}}) - \underset{\sim}{\Theta}] \mid \mathbf{C} \sim T^2_{mp}(n-m-q), \quad (9)$$

a Hotelling's T^2–variate with n-m-q degrees of freedom. Thus, given $\mathbf{C}$, we have:

$$\frac{(n-m-q)-mp+1}{mp(n-m-q)} T^2_{mp}(n-m-q) \sim F(mp, n-m-q-mp+1), \quad (10)$$

an F-variate with mp degrees of freedom in the numerator and n-m-q-mp+1 degrees of freedom in the denominator. Given $\mathbf{C}$, a confidence region for $\text{vec}\underset{\sim}{\Theta}$ with conditional coverage probability $1-\alpha$ is given by

$$\left\{\underset{\sim}{\Theta}: \frac{n-m-q-mp+1}{mp(n-m-q)} T^2_{mp}(n-m-q) \le F_{1-\alpha}(mp, n-m-q-mp+1)\right\} \quad (11)$$

with $T^2_{mp}(n-m-q)$ as defined on the left-hand side of (9) above. Since this confidence region has exact conditional coverage $1-\alpha$, it also has exact unconditional coverage $1-\alpha$.

5.4. Hypothesis Tests on the Metamodel

We consider hypotheses of the the form $H_0: \mathbf{G}\text{vec}\underset{\sim}{\Theta} = \mathbf{b}$, where $\mathbf{G}$ is a known s×mp matrix of full row rank and $\mathbf{b}$ is a known s×1 column vector. We have

$$[\mathbf{G}\,\text{vec}\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}}) - \mathbf{b}]'(\mathbf{GVG}')^{-1}[\mathbf{G}\,\text{vec}\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}}) - \mathbf{b}] \mid \mathbf{C} \sim T_s^2(n-m-q), \quad (12)$$

a Hotelling's T^2–variate with n-m-q degrees of freedom. A statistic for testing the null hypothesis H_0 is

$$\frac{(n-m-q)-s+1}{s(n-m-q)} T_s^2(n-m-q) \sim F(s, n-m-q-s+1), \quad (13)$$

an F-variate with s degrees of freedom in the numerator and n-m-q-s+1 degrees of freedom in the denominator.

6. RESULTS FOR THE APPLICATION

We performed the simulation experiment described in Section 3 using a discrete-event model written in the SLAM II simulation language (Pritsker (1986)). We executed the experiment on a CDC Dual Cyber 170/750 computer, where each run started with all users "thinking" at the terminals and each run stopped after 1500 seconds of simulated operation. To reduce the initialization bias on each run, we discarded the observations collected during the first 225 seconds of simulated time (15% of the run length). In general, one would probably use longer runs and truncate a

larger initial portion of each run. In any case, our objective was mainly to illustrate the methodology developed in this paper.

6.1. Estimates of the Metamodel Coefficients

Table 3 displays the results for the direct simulation estimator $\hat{\underset{\sim}{\Theta}}$ and for the controlled estimator $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$. Note that these results are based on a single repetition of the basic 81-point experimental design.

Table 3: Estimates of Metamodel Coefficients for the Interactive Computer System Simulation

Regression Coefficients	Direct Simulation Estimator $\hat{\underset{\sim}{\Theta}}$		Controlled Estimator $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$	
Constant	5.748	21.828	4.137	16.123
X_1	2.499	-.731	2.453	.599
X_2	-.834	2.691	-.608	2.232
X_3	1.926	1.614	.521	-.340
X_1^2	-.268	-.069	-.239	-.122
X_2^2	.267	-.342	.197	-.368
X_3^2	-.012	-.388	-.176	-.425
X_1X_2	-.007	-.141	-.023	-.166
X_1X_3	-.297	-.715	-.183	-.399
X_2X_3	-.065	.519	-.074	.367

6.2. A Test for Nonlinear Effects

Since we have postulated a bivariate model of response times that is a second-degree polynomial in the decision variables, we might want to test the statistical significance of the two-factor interactions and the quadratic effects in the hypothesized model. In this case, our null hypothesis is:

$$H_0 : \underset{\sim}{\Theta}_{i,j} = 0, \quad \text{for } i = 5, \ldots, 10 \text{ and } j = 1, 2.$$

In terms of the formulation of Section 5.4, we have:

$$\mathbf{G} = \begin{bmatrix} \mathbf{0}_{6\times4} & \mathbf{I}_6 & \mathbf{0}_{6\times4} & \mathbf{0}_{6\times6} \\ \mathbf{0}_{6\times4} & \mathbf{0}_{6\times6} & \mathbf{0}_{6\times4} & \mathbf{I}_6 \end{bmatrix},$$

with $s = \text{rank}(\mathbf{G}) = 12$, and $\mathbf{b} = \mathbf{0}_{12\times1}$. Evaluating the left-hand side of display (12), we obtain $T_s^2(n-m-q) = T_{12}^2(65) = 13.833$; and the F-ratio (13) has the value 0.9577. Since the computed F-ratio is based on 54 degrees of freedom in the denominator, we conclude that the nonlinear effects are negligible. Thus, we retain only the constant term and the first-order effects $\{X_1, X_2, X_3\}$ in (3).

6.3. A Confidence Region for the Final Metamodel

Now we obtain a confidence region for the final metamodel coefficient matrix $\underset{\sim}{\Theta}$. On the basis of the hypothesis test described in the previous section (which indicates that only the constant term and the linear effects in model (3) are nonzero), we seek to estimate the revised $\underset{\sim}{\Theta}$ using the 6 standardized control variables defined in Section 3 (namely, the standardized service times at service stations 3, 4, and 5 for classes 1 and 2). To avoid using the same data set for testing a model hypothesis and for estimating the revised model based on that test, we have independently repeated our 81-run simulation experiment to obtain the new point estimators given in Table 4.

Table 4: Final Estimates of Metamodel Coefficients for the Interactive Computer System Simulation

Regression Coefficients†	Direct Simulation Estimator $\hat{\underset{\sim}{\Theta}}$		Controlled Estimator $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$	
Constant	9.671	20.159	10.422	19.257
X_1	-.099	-.103	-.149	-.018
X_2	.070	-.126	.041	.110
X_3	-.370	.401	-.212	.380

†Only the constant term and the linear effects are included in the final metamodel.

A 90% confidence region for $\underset{\sim}{\Theta}$ can be obtained from equations (9), (10) and (11). As $F_{0.90}(8, 64) = 1.77$ and $(n-m-q-mp+1)/[mp(n-m-q)] = 64/568$, the 90% confidence ellipsoid centered at $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$ is

$$\left\{ \underset{\sim}{\Theta} : \text{vec}[\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}}) - \underset{\sim}{\Theta}]' \, \mathbf{V}^{-1} \, \text{vec}[\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}}) - \underset{\sim}{\Theta}] \leq 15.68 \right\},$$

where $\hat{\underset{\sim}{\Theta}}(\hat{\underset{\sim}{\Delta}})$ is given in Table 4, and

$$\mathbf{V}^{-1} =$$

$$\begin{bmatrix} 127.9 & 755.7 & 121.3 & -304.0 & -3.5 & -20.8 & -3.3 & 8.4 \\ & 4677.2 & 707.6 & -1792.2 & -20.8 & -129.0 & -19.5 & 49.4 \\ & & 217.9 & -289.1 & -3.3 & -19.5 & -6.0 & 8.0 \\ & & & 768.3 & 8.4 & 49.4 & 8.0 & -21.2 \\ & & & & 15.3 & 90.6 & 14.5 & -36.4 \\ & & & & & 560.7 & 84.8 & -214.8 \\ & & & & & & 26.1 & -34.7 \\ & & & & & & & 92.1 \end{bmatrix}$$

Perhaps a more easily interpretable confidence region results from computing *simultaneous* 90% confidence inter-

vals on all metamodel coefficients—that is, a confidence rectangle for $\underset{\sim}{\Theta}$. Such a confidence rectangle is given in Porta Nova (1985) and in Porta Nova and Wilson (1986).

7. CONCLUSIONS AND RECOMMENDATIONS

The minimum variance ratio (4), the variance ratio (7), the loss factor (8) and the confidence ellipsoid (11) all appear to be fairly robust to departures from normality in the simulation responses and/or controls. On the other hand, the validity of these quantitites seems to be highly sensitive to the degree of heterogeneity of the response variance across the points of the design. See Porta Nova (1985) and Porta Nova and Wilson (1986) for further discussion of these conclusions. A more extensive experimental performance evaluation is needed to support any truly general conclusions on these issues.

As an extension to the framework discussed in this paper, we recommend investigation of some resampling techniques (like jackknifing) to provide distribution-free results, for cases in which the assumptions regarding normality and/or homogeneity of variance are untenable. Other possible extensions include the analysis of nonlinear models and appropriate modifications of our procedures to exploit any extra information that is available about the covariance structure $\underset{\sim}{\Sigma}_Z$ of the target system.

REFERENCES

Arnold, S. F. (1981). *The Theory of Linear Models and Multivariate Analysis.* John Wiley & Sons, New York.

Kleijnen, J. P. C. (1974). *Statistical Techniques in Simulation, Part I.* Marcel Dekker, New York.

Lavenberg, S. S., Moeller, T. L., and Welch, P. D. (1982). Statistical results on control variables with application to queueing newtork simulation. *Operations Research* **30**, 182-202.

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

Pritsker, A. A. B. (1986). *Introduction to Simulation and SLAM II,* Third Edition. Halsted Press, New York.

Porta Nova, A. M. (1985). A generalized approach to variance reduction in discrete-event simulation using control variables. Unpublished Ph.D. Dissertation, Mechanical Engineering Department, University of Texas, Austin, Texas.

Porta Nova, A. M. and Wilson, J. R. (1986). Estimation of multiresponse simulation metamodels using control variates (to appear).

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

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics.* John Wiley & Sons, New York.

Venkatraman, S. (1983). Application of the control variate technique to multiresponse simulation output analysis. Unpublished M.S. Thesis, Mechanical Engineering Department, University of Texas, Austin, Texas.

Venkatraman, S. and Wilson, J. R. (1986). The efficiency of control variates in multiresponse simulation. *Operations Research Letters* **5**, 37-42.

Wilson, J. R. (1984). Variance reduction techniques for digital simulation. *American Journal of Mathematical and Management Sciences* **4**, 277-312.

AUTHORS' BIOGRAPHIES

ACÁCIO M. DE O. PORTA NOVA is an assistant professor in the Mathematics Department of the Superior Technical Institute (IST) of the Technical University of Lisbon. He received a B.S. in electrical engineering from IST in 1978, and he received a Ph.D. in operations research from The University of Texas at Austin in 1985. From 1978 to 1982 he was a lecturer at IST and a research associate in the Center for Urban and Regional Studies (CESUR) of the Technical University of Lisbon. From 1979 to 1982 he was also a systems analyst in the Computation Center of the Universities of Lisbon. Since 1986 he has been a researcher in CESUR. His research interests include simulation output analysis, variance reduction techniques, numerical analysis, and software engineering.

Acácio M. de O. Porta Nova
Departamento de Matemática
Instituto Superior Técnico
Avenida Rovisco Pais

1096 Lisboa CODEX, Portugal
011-351-1-809580

JAMES R. WILSON is an associate professor in the School of Industrial Engineering at Purdue University. He received a B.A. in mathematics from Rice University in 1970, and M.S. and Ph.D. degrees in industrial engineering from Purdue University in 1977 and 1979 respectively. He has been involved in various simulation studies while working as a research analyst for the Houston Lighting & Power Company (1970-72) and while serving as a U.S. Army officer (1972-75). From 1979 to 1984, he was an assistant professor in the Mechanical Engineering Department of The University of Texas at Austin. His current research interests include simulation output analysis, variance reduction techniques, ranking-and-selection procedures, and stopping rules. He is a member of ACM, IIE, ORSA, SCS, and TIMS.

James R. Wilson
School of Industrial Engineering
Purdue University
West Lafayette, IN 47907, U.S.A.
(317) 494-5408