Proceedings of the 1988 Winter Simulation Conference
M. Abrams, P. Haigh, and J. Comfort (eds.)

A comparison of confidence region estimators for multivariate simulation output

John M. Charnes
W. David Kelton
Department of Management Sciences
Curtis L. Carlson School of Management
University of Minnesota
Minneapolis, MN 55455, U.S.A.

ABSTRACT

Three previously proposed methods for constructing joint confidence regions on the mean of multivariate simulation output are described and tested using data generated by Gaussian vector autoregressive moving average models. The methods that use an estimate of the variance-covariance matrix of the data are found to yield regions with lower volumes than the method that does not use an estimate of the variance-covariance matrix. The experimental design included the factors run length, dimension, autocorrelation and cross correlation.

1. INTRODUCTION

Most research in simulation output analysis has been concerned primarily with analysis of a univariate sequence of observations; however, the interesting practical problems seldom are such that they can be summarized by a single parameter (Seila 1984). For instance, a manufacturing environment may be modeled as a network of queues in which the activity in each queue somehow affects the other queues. Thus, previous research has effectively ignored the interrelationships among variables that may be of most interest to practitioners.

A naive method of analyzing multivariate simulation output such as that referred to above would be to construct individual confidence intervals on the mean of a measure of interest for each queue, all at the same nominal level of confidence. However, to do this one would not only be ignoring the possible interrelationships among the queues but would also be ignoring the so-called "multiple comparisons problem" stemming from the fact that the overall level of confidence one may have in all of the intervals' containing their respective true means is lower than the nominal level stated for each interval. The more intervals constructed, the lower will be the overall level of confidence.

Previous researchers have addressed the multiple comparisons problem with regard to simulation output analysis (Seila 1984, Chen and Seila 1987, Kabaila and Nelson 1985), but the multivariate techniques that have been proposed to account for the interrelationships among output processes have only been justified asymptotically so that their performance with the moderate amounts of data generally available in practice remains largely untested.

This paper is intended to present the results of a controlled simulation study designed to compare three multivariate output analytic methods proposed thus far in the literature. Section 2 describes the Bonferroni method, section 3 describes the multivariate batch means method, section 4 describes the spectral analysis method and section 5 gives the results of the comparison and discusses the implications of the study.

2. BONFERRONI METHOD

The use of Bonferroni intervals for simultaneous inference in multivariate statistical analysis is well known (see Johnson and Wichern 1982 or Bickel and Doksum 1977). The method is based on the Bonferroni inequality, which gives a lower bound on the probability of the joint occurrence of a finite number of events. In particular, let

C_i denote a confidence statement about the mean value of a stationary time series. If

$$P[C_i \text{ is true}] = 1 - \alpha_i \text{ for } i = 1,\ldots,d$$

then

$$P[\text{all } C_i \text{ true}] \geq 1 - \alpha_1 + \alpha_2 + \ldots + \alpha_d).$$

Thus, for instance, if each one of d confidence intervals is constructed at the $1 - \alpha/d$ level using a valid method, one may have an overall level of confidence of at least $1 - \alpha$ that all d parameters lie in the confidence region formed by the d-dimensional hyperbox defined by the d confidence intervals.

The use of Bonferroni intervals may be advised when d is small (Johnson and Wichern 1982) but, for large values of d, Bonferroni intervals may lead to excessively wide (relative to the amount of data available) intervals for each of the parameters unless there is very little underlying variability. Moreover, even when the time series for which the simultaneous intervals are being constructed are independent, the Bonferroni intervals will be conservative. Furthermore, if the time series are interrelated, the use of Bonferroni intervals may lead to "excessively large" confidence regions (Kabaila and Nelson 1985).

In using the Bonferroni method, one must make a choice as to which univariate method one will use for constructing the individual intervals. The univariate batch means method using the Von Neumann ratio for batch size selection (Fishman (1978b) was chosen here because of the preliminary nature of the study and the widespread use of the batch means method.

3. MULTIVARIATE BATCH MEANS METHOD

The Multivariate Batch Means method of constructing confidence regions is a generalization of the univariate batch means method of constructing confidence intervals (Kabaila and Nelson 1985). Chen and Seila (1987) give details for the application of the Multivariate Batch Means method to steady-state simulation output.

Consider a simulation operating in steady-state that produces a sequence of vector-valued observations $\mathbf{X}_i$, i = 1, ..., n where each vector $\mathbf{X}_i$ has d components: $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^T$. ($\mathbf{x}^T$ denotes the transpose of the vector $\mathbf{x}$.) The simulation analyst wishes to estimate the d-dimensional mean vector $\mu = E(\mathbf{X}_i) = (E(\mathbf{X}_{i1}), \ldots, E(\mathbf{X}_{id})^T$ with a joint confidence region. The multivariate batch means method calls for dividing the sequence of output vectors into k batches of m (vector) observations each (where n = mk) and computing the sample mean vector for each batch. If m is chosen large enough, then the k vectors of batch means can be treated as if they are uncorrelated vectors of observations (Chen and Seila 1987). Then a confidence region on the mean vector, μ, is constructed by first computing an estimate, $\mathbf{S}$, of the variance-covariance matrix in a standard multivariate statistical analytic fashion. An approximate $100(1-\alpha)\%$ confidence region for μ is then given by the boundary and interior of the ellipsoid,

$$\{\theta: k(\bar{\mathbf{X}} - \theta)^T \mathbf{S}^{-1} (\bar{\mathbf{X}} - \theta) \leq \frac{d(k-1)}{k-d} F_{\alpha;d,k-d}\}$$

where $F_{\alpha;d,k-d}$ is the $100(1-\alpha)\%$ quantile of the F-distribution with d numerator and k-d denominator degrees of freedom.

Chen and Seila (1987) give a sequential procedure for choosing m in which the number of observations per batch increases until the sample first-order correlation coefficient for the sequence of vector batch means does not appear to differ significantly from zero or the number of batches is less than eight, whichever comes first. This procedure is the multivariate analog of the approach taken by Fishman (1978a).

They test their method empirically by using a queuing model consisting of Poisson arrivals to two servers with independent

service times to generate 3-dimensional vector time series observations on waiting time in queue 1; waiting time in queue 2; and the total time in system. The coverage rates for a 97.5% confidence region ranged from a low of 62% when the sample size was 2000 observations to 73% when the sample size was 5000 observations for the model with a high (0.9) traffic intensity. For a moderate (0.75) traffic intensity, the range of coverage was from 80% (2000 observations) to 91% (5000 observations), while the low (0.5) traffic intensity results ranged from 94% coverage (2000 observations) to 97% coverage (5000 observations). Given the dearth of research in multivariate simulation analysis to date, it is hard to assess the performance of the proposed technique relative to the performance of any competing techniques.

4. SPECTRAL ANALYSIS

Kabaila and Nelson (1985) proposed a method of setting up a confidence region for the mean of a multivariate time series which was motivated by the analysis of simulation models used in an attempt to identify the earth's mean atmospheric response to external forcing. Such models produce multivariate output (e.g., temperatures at different geographic locations) observed at equally-spaced points in time. The statistical problem of estimating the mean of a vector process is fundamentally the same for atmospheric as well as discrete-event simulation models.

Just as in the univariate case, for a strictly stationary and phi-mixing d-dimensional process $\{\mathbf{X}_t,\ t = 1, 2, \ldots, n\}$ with an absolutely continuous spectrum, it can be shown that

$$\sqrt{n}(\bar{\mathbf{X}} - \mu) \xrightarrow{D} N(\mathbf{0}, 2\pi\mathbf{f}(0))$$

where $\xrightarrow{D}$ denotes convergence in distribution and $\mathbf{f}(0)$ is the spectral density matrix of $\mathbf{X}(t)$ evaluated at frequency zero. Kabaila and Nelson's procedure is based on this result. Following Hannan (1970), they suggest the following estimator of $2\pi\mathbf{f}(0)$

$$\mathbf{S}_n = \frac{1}{m} \sum_{u=1}^{m} \{\delta(u)\delta(u)^T + \phi(u)\phi(u)^T\}$$

where $(2m \geq d)$ and

$$\delta(u) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{X}(t) \cos(t\Gamma_u)$$

$$\phi(u) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{X}(t) \sin(t\Gamma_u)$$

$$\Gamma_u = \frac{2\pi u}{n}$$

m is a value whose choice affects the expected volume of the subsequent confidence region and the difference between the nominal confidence of the region and its actual confidence. Kabaila and Nelson give vague guidelines for the optimal choice of the value of m when the processes are Gaussian and call for a simulation study of data-aided procedures for choosing m. We somewhat arbitrarily chose the values of 20 and 100 for m in an initial attempt to get a feel for what setting may be appropriate for this parameter.

After finding $\mathbf{S}_n$, Kabaila and Nelson's proposed technique calls for constructing the following $100(1-\alpha)\%$ confidence region for μ:

$$\{\theta:\ (\bar{\mathbf{X}} - \theta)^T \mathbf{S}_n^{-1} (\bar{\mathbf{X}} - \theta) \leq \frac{2md}{n(2m-d+1)} F_{\alpha;d,2m-d+1}\}$$

where $\mathbf{S}_n^{-1}$ is the matrix inverse of $\mathbf{S}_n$. This confidence region is the interior and boundary of an ellipsoid in d-dimensional space centered at $\mathbf{X}$ with size and shape depending on $\mathbf{S}_n$ and α.

Spectral-analytic methods of univariate confidence interval construction are known to be somewhat complicated and expensive to implement. This multivariate generalization of the univariate methods shares those same characteristics.

5. RESULTS OF COMPARISON

In the following discussion and tables, the methods will be referred to by the following abbreviations: BBM - Bonferroni method using batch means as the univariate confidence interval construction method; MBM - Chen and Seila's (1987) multivariate batch means method; S20 - Kabaila and Nelson's (1985) spectral analysis method with m = 20; and, S100 - Kabaila and Nelson's (1985) spectral analysis method with m = 100.

The data that were analyzed by each of the three methods were generated from Gaussian vector autoregressive moving-average (VARMA) models. Four factors were varied to assess their effects on the volume and coverage obtained by each method. Further, the methods were all applied to each vector output process so that a comparison of the methods in terms of volume and coverage could be made.

The four factors dimension, run length, autocorrelation, and cross correlation were varied between two levels as shown in Table 1. Appendix A gives the specification of the generating models and shows the autocorrelation matrices at lags 0 and 1 representing low and high levels of autocorrelation and cross correlation for each generating model.

Table 2 shows that the S20 method seemed to yield the best coverage with an average coverage rate of 0.95 which is equal to the nominal level of confidence specified for each of the four methods. However, all methods gave coverage rates which were quite close to the 0.95 nominal level, a fact which is likely due to the use of Gaussian disturbances in the VARMA data-generating models. A more interesting comparison is to look at the volumes of the regions resulting from the four methods.

The MBM method yielded consistently lower average volumes at each design point than any other method, though the difference between MBM and S100 may not be practically important. Table 2 does show quite clearly that the ellipsoidal confidence regions given by MBM, S100, and S20 are significantly smaller than the hyperbox region given by the conservative BBM method. This indicates that there is some benefit from using the information provided by the variance-covariance matrix on the dependencies inherent in the data. This information on dependencies is not used by the Bonferroni method.

The execution times shown in Table 3 are the average central processor time used in executing each method on a Cray-2 computer scaled so that the BBM method's time is 1.0. (The Cray-2 required 12.4 minutes of central

Table 1: Design Point Coding Scheme

Design Point	Dimension (d)	Run Length (N)	Autocorrelation	Cross correlation
1	2	5000	Low	High
2	2	5000	High	Low
3	2	10000	Low	Low
4	2	10000	High	High
5	4	5000	Low	Low
6	4	5000	High	High
7	4	10000	Low	High
8	4	10000	High	Low

Table 2: Confidence Region Coverage and Volume (x 10,000)

Design Point	Coverage BBM	MBM	S20	S100	Volume (x 10,000) BBM	MBM	S20	S100
1	0.92	0.88	0.94	0.90	44.638	37.973	41.211	38.210
2	0.92	0.96	0.94	0.96	41.710	37.542	39.877	37.853
3	0.90	0.94	0.96	0.94	21.252	18.764	20.034	19.079
4	0.94	0.96	0.94	0.96	22.127	18.907	20.563	19.159
5	0.98	0.94	1.00	0.96	0.564	0.244	0.297	0.250
6	0.94	0.92	0.92	0.92	0.584	0.237	0.306	0.240
7	0.94	0.92	0.96	0.92	0.151	0.059	0.075	0.061
8	0.94	0.92	0.94	0.90	0.144	0.059	0.074	0.061
Average	0.935	0.930	0.950	0.933				

processor time to produce 50 sets of observations for each method at each design point.) Table 3 indicates that the spectral-analytic methods' reputations for being expensive are not in doubt here. The S100 method, while approaching the same coverage rates and volume as MBM, took nearly 100 times as long to execute. It is not clear whether this ratio would change if the same comparison were done on a computer that does not "vectorize" the code as does the Cray-2.

Table 3: Relative Execution Times

Method	Relative Time
BBM	1.0
MBM	5.8
S20	121.9
S100	569.0

The main effects of the experimental factors shown in Tables 4 and 5 are somewhat surprising in what they do not show. The coverage does not seem to show any effect from varying any factor and the only significant main effects in volume seems to come about from varying the Run Length factor, which is not too surprising. The effect from the cross correlation factor on BBM for the 2-dimensional case (1.90) and the effect from the cross correlation factor on S100 for the 4-dimensional case (-0.005) are more than two standard errors from zero, but these may well be spurious in light of all the numbers presented in Tables 4 and 5.

In conclusion, it seems that the MBM method is the clear winner for the methods tested here on Gaussian VARMA-generated data. While the S100 method comes close to producing the same coverage and volume as MBM, the much longer execution time for S100 is discouraging. Still, it may behoove one to investigate these techniques further, given the significantly lower volume of the MBM region compared to the BBM region.

An interesting extension of this comparison would be to look at the results of testing queuing and inventory simulation data which was not Gaussian. Further refinements could include an improvement in the method of picking m in the spectral analytic methods and testing for independence of the batch mean vectors in the MBM method. It may also be interesting to look at alternative methods (such as the autoregressive method) of estimating $2\pi f(0)$ in the spectral analysis method.

Table 4: Factor Main Effects

2-Dimensional Output Vector

(Standard Errors are in parentheses)

	Coverage				10,000 x Volume			
Factor	BBM	MBM	S20	S100	BBM	MBM	S20	S100
Run Length	0.00	0.03	0.01	0.02	-21.48*	-18.92*	-20.25*	-18.91*
	(0.04)	(0.03)	(0.03)	(0.03)	(0.99)	(0.24)	(0.84)	(0.30)
Autocorr.	0.02	0.05	-0.01	0.04	-1.03	-0.14	-0.40	-0.14
	(0.04)	(0.04)	(0.04)	(0.04)	(1.03)	(0.26)	(0.87)	(0.35)
Cross corr.	0.02	-0.03	-0.01	-0.02	1.90*	0.29	0.93	0.22
	(0.03)	(0.03)	(0.03)	(0.03)	(0.90)	(0.26)	(0.80)	(0.28)

* Asterisk denotes an effect which is more than two standard errors from zero.

Table 5: Factor Main Effects

4-Dimensional Output Vector

(Standard Errors are in parentheses)

	Coverage				10,000 x Volume			
Factor	BBM	MBM	S20	S100	BBM	MBM	S20	S100
Run Length	-0.02	-0.01	-0.01	-0.03	-0.43*	-0.18*	-0.23*	-0.18*
	(0.03)	(0.03)	(0.03)	(0.04)	(0.02)	(0.001)	(0.008)	(0.002)
Autocorr.	-0.02	-0.01	-0.05	-0.03	0.01	-0.004	0.004	-0.005
	(0.03)	(0.04)	(0.03)	(0.04)	(0.02)	(0.002)	(0.008)	(0.003)
Cross corr.	-0.02	-0.01	-0.03	-0.01	0.01	-0.003	0.005	-0.005*
	(0.03)	(0.04)	(0.03)	(0.04)	(0.01)	(0.002)	(0.007)	(0.002)

* Asterisk denotes an effect which is more than two standard errors from zero.

APPENDIX: MODEL SPECIFICATION

The data for this study were generated from either a VARMA(1,0) or a VARMA(1,1) model, depending on the particular combination of the autocorrelation and cross correlation factor levels. This appendix shows the autocorrelation matrices at lags 0 and 1 corresponding to the combinations of factor levels used in the experiment.

2-Dimensional Data

Low autocorrelation, Low cross correlation

Lag 0 autocorrelation	Lag 1 autocorrelation
1.00 0.08	0.10 0.01
0.08 1.00	0.08 0.11

Low autocorrelation, High cross correlation

Lag 0 autocorrelation	Lag 1 autocorrelation
1.00 0.99	0.10 0.10
0.99 1.00	0.17 0.11

High autocorrelation, Low cross correlation

Lag 0 autocorrelation	Lag 1 autocorrelation
1.00 0.04	0.71 0.03
0.04 1.00	0.03 0.71

High autocorrelation, High cross correlation

Lag 0 autocorrelation	Lag 1 autocorrelation
1.00 0.92	0.75 0.68
0.92 1.00	0.87 0.88

4-Dimensional Data

Low autocorrelation, Low cross correlation

Lag 0 autocorrelation	Lag 1 autocorrelation
1.00 0.06 0.08 0.08	0.10 0.01 0.01 0.01
0.06 1.00 0.09 0.09	0.11 0.11 0.02 0.02
0.08 0.09 1.00 0.07	0.08 0.08 0.11 0.02
0.08 0.09 0.07 1.00	0.09 0.09 0.12 0.12

Low autocorrelation, High cross correlation

Lag 0 autocorrelation	Lag 1 autocorrelation
1.00 0.70 0.98 0.96	0.10 0.07 0.10 0.01
0.70 1.00 0.99 0.98	0.17 0.17 0.20 0.19
0.98 0.99 1.00 0.72	0.22 0.22 0.24 0.21
0.96 0.98 0.72 1.00	0.31 0.31 0.30 0.30

High autocorrelation, Low cross correlation

Lag 0 autocorrelation	Lag 1 autocorrelation
1.00 0.05 0.07 0.07	0.71 0.04 0.05 0.05
0.05 1.00 0.07 0.07	0.04 0.71 0.05 0.05
0.07 0.07 1.00 0.05	0.05 0.05 0.71 0.04
0.07 0.07 0.05 1.00	0.05 0.05 0.04 0.71

High autocorrelation, High cross correlation

Lag 0 autocorrelation	Lag 1 autocorrelation
1.00 0.87 0.70 0.52	0.70 0.58 0.46 0.32
0.87 1.00 0.94 0.81	0.89 0.88 0.78 0.62
0.70 0.94 1.00 0.95	0.84 0.97 0.94 0.83
0.52 0.81 0.95 1.00	0.72 0.93 0.99 0.96

REFERENCES

Bickel, P.J. and Doksum, K.A. (1977). Mathematical Statistics, Holden-Day.

Chen, D.R. and A.F. Seila (1987). "Multivariate inference in stationary simulation using batch means", Proceedings of the 1987 Winter Simulation Conference, A. Thesen, H. Grant, W. David Kelton (eds.), pp. 302-304.

Fishman, G.S. (1978a). Principles of Discrete Event Simulation. Wiley: New York.

Fishman, G.S. (1978b). Grouping observations in digital simulation. Management Science 24(5), 510-521.

Hannan, E.J. (1970). Multiple Time Series, Wiley: New York.

Johnson, R.A. and Wichern, D.W. (1982). Applied Multivariate Statistical Analysis. Prentice-Hall: Englewood Cliffs, New Jersey.

Kabaila P. and G. Nelson (1985). "On confidence regions for the mean of a multivariate time series", Communications in Statistics - Theory and Methods, 14(3), pp. 735-753.

Seila, A.F. (1984). "Multivariate output analysis in discrete event simulation: The state of the art", Working Paper No. 84-151, College of Business Administration, The University of Georgia.

AUTHOR'S BIOGRAPHIES

John M. Charnes holds the B. Civil Eng. and M.B.A. degrees from the University of Minnesota and is currently a Ph.D. candidate in the Department of Management Sciences at the University of Minnesota.

John M. Charnes
Department of Management Sciences
Carlson School of Management
University of Minnesota
Minneapolis, Minnesota 55455
jcharnes@umnacvx.bitnet

W. David Kelton is an Associate Professor in the Management Sciences Department at the Carlson School of Management, University of Minnesota, in Minneapolis. He received a B.A. in Mathematics from the University of Wisconsin-Madison, an M.S. in Mathematics from Ohio University, and M.S. and Ph.D. degrees in Industrial Engineering from the University of Wisconsin-Madison. His research interests are in simulation methodology, stochastic modeling and estimation, and quality control. He is a member of ORSA, TIMS, ASA, and SCS. He served as Program Chair for the 1987 Winter Simulation Conference.

W. David Kelton
Department of Management Sciences
Carlson School of Management
University of Minnesota
Minneapolis, Minnesota 55455
612/624-8503
dkelton@umnacvx.bitnet