Technical Note

Multiple Comparisons with the Best for Steady-State Simulation

MINGJIAN YUAN
National Yunlin Institute of Technology and Management
and
BARRY L. NELSON
The Ohio State University

Multiple-comparison procedures are useful for comparing the performance of competing systems via simulation. In this paper we extend a particular multiple-comparison procedure, multiple comparisons with the best, to steady-state simulation by using an autoregressive-output-analysis method.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]; I.6.1 [**Simulation and Modeling**]: Simulation Theory

General Terms: Experimentation, Theory

Additional Key Words and Phrases: Computer simulation, multiple-comparison procedures, statistics, steady-state simulation, stochastic processes, times-series analysis

1. INTRODUCTION

An important use of simulation analysis is to compare and select the best from among a number of competing systems. This paper presents a procedure that can be applied when there are a finite number of systems and when comparisons are based on long-run expected performance.

In [14] we updated a method for constructing a confidence interval for the long-run expected performance of a *single* system based on an old idea:

This work was supported by National Science Foundation grant DDM-8922721.
Authors' addresses: M. Yuan, Department of Industrial Management, National Yunlin Institute of Technology and Management, Yunlin, Taiwan; and B. L. Nelson, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, OH 43210.

autoregressive modeling. Here we extend this method to the problem of *multiple comparisons*, that is, constructing simultaneous confidence intervals on differences in the expected performance of two or more systems.

In the next section, we review multiple comparisons and the results of [14]. Section 3 presents the new procedure and proves that it is asymptotically valid. In Section 4 we empirically evaluate the small-sample properties of the procedure. Conclusions and recommendations are given in Section 5.

2. BACKGROUND

There are many practical decision problems in which the number of competing systems is finite. Allocating a finite amount of buffer space between workstations in a production line is an example. Let $\theta_1, \theta_2, \ldots, \theta_k$ be the performance parameters associated with k competing systems, and let $\theta_{[k]} \geq \theta_{[k-1]} \geq \cdots \geq \theta_{[1]}$ be the ranked system performance parameters, where $[i]$ denotes the unknown index of the ith ranked system. Suppose that a larger performance parameter is better and that we want to find the best system.

Hsu and Nelson [11] showed that multiple comparisons with the best (MCB) can be applied to such problems. MCB constructs simultaneous confidence intervals for the parameter set $\theta_i - \max_{\ell \neq i} \theta_\ell$, for $i = 1, 2, \ldots, k$. If there is a unique best system, then only $\theta_{[k]} - \max_{\ell \neq [k]} \theta_\ell$ is positive, while all of the other parameters are negative. For systems other than the $[k]$th, $\theta_i - \max_{\ell \neq i} \theta_\ell$ is the difference between the performance of the ith system and the best system. MCB confidence intervals establish whether the sample best system can be declared to be the true best system, based on the data, and also how far the performance of each system might be from the performance of the best system.

MCB inference is frequently more conclusive than, for example, all-pairwise multiple comparisons ($\theta_i - \theta_\ell$ for all $i \neq \ell$). This is because MCB forms k simultaneous confidence intervals, whereas all-pairwise multiple comparisons require $k(k-1)/2$ intervals to be simultaneously correct; the more intervals that must be simultaneously correct, the wider the intervals tend to be.

Figure 1 illustrates two possible sets of MCB intervals. The dots represent the unknown parameters $\theta_i - \max_{\ell \neq i} \theta_\ell$, for $i = 1, 2, \ldots, 5$, and the vertical lines through the dots represent the associated confidence intervals. Notice that MCB intervals are constrained to contain 0 or have 0 as one endpoint. In Figure 1a all five intervals are correct, since they cover the parameters; MCB controls the probability of this event, which we call *correct* inference. The inference is also *conclusive*, since system 3 can be identified as the best (the lower endpoint of its interval indicates that the difference between system 3 and the best of the other systems is greater than or equal to 0). In Figure 1b the intervals are correct, but not conclusive; since two intervals contain 0, neither of the corresponding systems can be inferred to be different from the best. However, the lower endpoints of the intervals bound (with prespecified confidence level) how far the performance of these systems is from the performance of the best system.

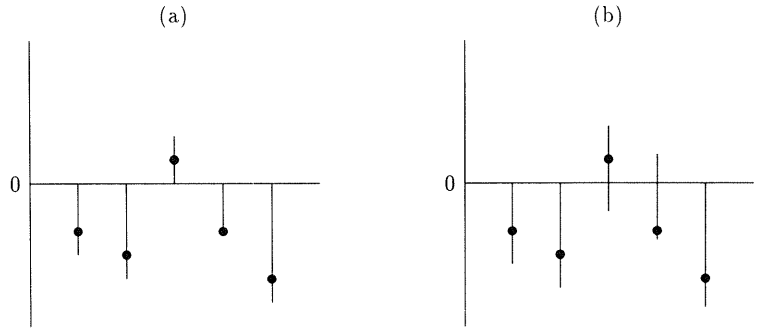

Fig. 1. MCB confidence intervals for $\theta_i - \max_{\ell \neq i} \theta_\ell$, for $i = 1, 2, \ldots, 5$.

Let Y_{ij} denote the jth observation of system performance from alternative i, for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, k$. Throughout this paper we assume that, for each system i, $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ is a stationary stochastic process, typically from a single replication of a steady-state simulation, and that $E[Y_{ij}] = \theta_i$. In addition, we only consider independent experiments; that is, the observations across systems are independent.

Let $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$ be the point estimator for θ_i. The following are sufficient conditions for forming MCB intervals [12]:

(1) The joint distribution of the sample-mean vector $[\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k]'$ is

$$\begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \vdots \\ \bar{Y}_k \end{bmatrix} \sim \mathrm{N}_k \left(\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{bmatrix}, \sigma^2 \begin{bmatrix} \delta_1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_k \end{bmatrix} \right), \tag{1}$$

where N_k stands for the multivariate normal distribution of dimension k and where the δ_i's are known constants.

(2) An estimator S^2 of σ^2 is available that is independent of the sample-mean vector, and the distribution of S/σ is known.

For example, suppose the data satisfy a one-way analysis-of-variance model

$$Y_{ij} = \theta_i + \epsilon_{ij},$$

for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, k$, where the ϵ_{ij} are i.i.d. $\mathrm{N}(0, \sigma^2)$ random variables. Then both conditions are satisfied by letting S^2 be the usual pooled variance estimator of σ^2 and by letting $\delta_i = 1/n_i$. Unfortunately, this model is not appropriate for steady-state simulation when the outputs from within each system are dependent.

We now display the MCB intervals associated with the sufficient conditions

enumerated above. Let $\mathbf{n} = [n_1, n_2, \ldots, n_k]$, and let

$$U_i(\mathbf{n}) = \left(\min_{\ell \neq i} \left\{ \overline{Y}_i - \overline{Y}_\ell + d^i_{1-\alpha,f} S \sqrt{\delta_i + \delta_\ell} \right\} \right)^+,$$

$$\mathscr{C} = \left\{ i: \min_{\ell \neq i} \left\{ \overline{Y}_i - \overline{Y}_\ell + d^i_{1-\alpha,f} S \sqrt{\delta_i + \delta_\ell} \right\} > 0 \right\},$$

$$L_i(\mathbf{n}) = \begin{cases} 0, & \text{if } \mathscr{C} = \{i\}, \\ -\left(\min_{\ell \in \mathscr{C}, \ell \neq i} \left\{ \overline{Y}_i - \overline{Y}_\ell - d^\ell_{1-\alpha,f} S \sqrt{\delta_\ell + \delta_i} \right\} \right)^-, & \text{otherwise,} \end{cases}$$

where $a^+ = \max\{0, a\}$, $a^- = -\min\{0, a\}$, and $d^i_{1-\alpha,f}$ is the critical value such that

$$\Pr\left\{ \overline{Y}_i - \theta_i \geq \overline{Y}_\ell - \theta_\ell - d^i_{1-\alpha,f} S \sqrt{\delta_i + \delta_\ell}, \forall \ell \neq i \right\} = 1 - \alpha. \tag{2}$$

THEOREM 2.1 [9, 10]. *If the sufficient conditions* (1) *and* (2) *are satisfied, then*

$$\{[L_1(\mathbf{n}), U_1(\mathbf{n})], [L_2(\mathbf{n}), U_2(\mathbf{n})], \ldots, [L_k(\mathbf{n}), U_k(\mathbf{n})]\}$$

form a set of $(1 - \alpha)100$ *percent simultaneous confidence intervals for* $\theta_i - \max_{\ell \neq i} \theta_\ell$, $i = 1, 2, \ldots, k$.

For the situations considered in this paper, $d^i_{1-\alpha,f}$ is the $1 - \alpha$ quantile of the maximum of a $(k - 1)$-dimensional multivariate t random variable with degrees of freedom f and correlation matrix:

$$\mathbf{R}_i = \begin{bmatrix} 1 & \lambda^{(i)}_1 \lambda^{(i)}_2 & \cdots & \lambda^{(i)}_1 \lambda^{(i)}_{k-1} \\ \lambda^{(i)}_2 \lambda^{(i)}_1 & 1 & \cdots & \lambda^{(i)}_2 \lambda^{(i)}_{k-1} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda^{(i)}_{k-1} \lambda^{(i)}_1 & \lambda^{(i)}_{k-1} \lambda^{(i)}_2 & \cdots & 1 \end{bmatrix}, \tag{3}$$

where

$$1/\lambda^{(i)}_\ell = \begin{cases} \sqrt{1 + \delta_\ell/\delta_i}, & \text{if } \ell < i, \\ \sqrt{1 + \delta_{\ell+1}/\delta_i}, & \text{if } \ell \geq i. \end{cases} \tag{4}$$

See [8, pp. 374–375] for a definition of the multivariate t distribution and for the computation of critical values like $d^i_{1-\alpha,f}$.

To extend MCB to steady-state simulation, assume that the output process Y_{ij}, $j = 1, 2, \ldots, n_i$, is a stationary AR(p_i) process defined by the model

$$\begin{aligned} Y_{ij} &= \theta_i + \sum_{m=1}^{p_i} \phi_{im}(Y_{i,j-m} - \theta_i) + \epsilon_{ij} \\ &= \phi_{i0} + \sum_{m=1}^{p_i} \phi_{im} Y_{i,j-m} + \epsilon_{ij}, \end{aligned} \tag{5}$$

where the ϵ_{ij} are i.i.d. mean 0, variance σ_i^2 random variables, and $\phi_{i0} = \theta_i(1 - \sum_{m=1}^{p_i} \phi_{im})$. For system i, p_i and ϕ_{im} are the autoregressive order and coefficients, respectively, and σ_i^2 is the residual variance.

Under mild conditions, an AR model can match the autocovariance structure of an output process to any finite number of lags. Thus, a procedure that assumes that the output process is AR should provide a good approximation. The theoretical basis for the AR-confidence-interval procedure we employ is given in [14], but we briefly review it here. For ease of presentation, we temporarily drop the subscript i, denoting system.

(1) *AR order identification.* There are a number of procedures and criteria for AR order identification. They include Hannan's F test [5], Akaike's information criterion [1], Box and Jenkins's recursive procedure [2], Gray, Kelly, and McIntire's "D" statistics [4], and Rissanen's predictive-least-squares (PLS) criterion [13]. We use Rissanen's PLS criterion because it provides a strongly consistent order estimator; strong consistency is a property we exploit to prove the asymptotic validity of our MCB procedure.

Let

$$\mathrm{PLS}_h = \frac{\sum_{j=2h+2}^{n} e_j^2(h)}{n - 2h - 1},$$

where $e_j(h) = Y_j - \hat{Y}_j(h)$ is the honest prediction error for Y_j, assuming that the AR order is h, and where $\hat{Y}_j(h)$ is the predictor of Y_j, based on the AR(h) model estimated from $Y_1, Y_2, \ldots, Y_{j-1}$. The PLS criterion estimates the order to be the value $\hat{p}$ such that

$$\mathrm{PLS}_{\hat{p}} = \min_{h \in \mathscr{O}} \mathrm{PLS}_h,$$

where $\mathscr{O}$ is the set of possible orders. We assume that $\mathscr{O}$ contains the true order. An important result is the following:

THEOREM 2.2 [6, 7]. *If model* (5) *pertains, then* $\hat{p} \to p$ *with probability* 1 *as* $n \to \infty$.

(2) Use the conditional-least-squares estimators (CLSE) to estimate the coefficients. Suppose $\hat{p}$ is the PLS order estimator. Conditional on the first $\hat{p}$ observations, we can express the relationship between the remaining $n - \hat{p}$ observations as a linear model in the unknown AR coefficients. The AR coefficients are then estimated via least squares.

Let $\hat{\boldsymbol{\phi}}(n, \hat{p}) = [\hat{\phi}_0(n, \hat{p}), \hat{\phi}_1(n, \hat{p}), \ldots, \hat{\phi}_{\hat{p}}(n, \hat{p})]'$ be the CLSE in conjunction with the PLS order estimator. Then we have the following result:

THEOREM 2.3 [14]. *If model* (5) *pertains, then* $\hat{\boldsymbol{\phi}}(n, \hat{p}) \xrightarrow{\mathscr{D}} \boldsymbol{\phi}$ *as* $n \to \infty$, *where* $\boldsymbol{\phi} = [\phi_0, \phi_1, \ldots, \phi_p]'$ *is the true coefficient vector, and* $\xrightarrow{\mathscr{D}}$ *denotes convergence in probability.*

From here on we drop the argument $(n, \hat{p})$ from $\hat{\phi}_m(n, \hat{p})$ where there will be no confusion.

(3) Estimate the residual variance, σ^2, by

$$\hat{\sigma}^2 = \frac{\sum_{j=\hat{p}+1}^{n}\left(Y_j - \hat{\phi}_0 - \sum_{m=1}^{\hat{p}} \hat{\phi}_m Y_{j-m}\right)^2}{n - \hat{p}}. \tag{6}$$

(4) Estimate the variance of $\bar{Y}$ by

$$\widehat{\mathrm{Var}}(\bar{Y}) = \frac{\hat{\sigma}^2}{n\left(1 - \sum_{m=1}^{\hat{p}} \hat{\phi}_m\right)^2}.$$

THEOREM 2.4 [14]. *If model* (5) *pertains, then* $\hat{\sigma}^2 \xrightarrow{\mathcal{P}} \sigma^2$, $n\widehat{\mathrm{Var}}(\bar{Y}) \xrightarrow{\mathcal{P}} \sigma^2/(1 - \sum_{m=1}^{p} \phi_m)^2$, *and*

$$\frac{\bar{Y} - \theta}{\sqrt{\widehat{\mathrm{Var}}(\bar{Y})}} \xrightarrow{\mathcal{D}} \mathrm{N}(0,1)$$

as $n \to \infty$, *where* $\xrightarrow{\mathcal{D}}$ *denotes convergence in distribution.*

(5) Approximate the distribution of the sample variance, $\widehat{\mathrm{Var}}(\bar{Y})$, as a constant times an χ^2 random variable. If the sample size is very large, take the degrees of freedom to be n. If the sample size is small, use Fishman's approximation [3] or use the equivalent sample size, $\max\{1, \lfloor n(1 - \sum_{m=1}^{\hat{p}} \hat{\phi}_m)^2 \rfloor\}$.

In the next section, we extend the results of [14] for univariate confidence intervals to multiple comparisons using MCB.

3. MCB FOR STEADY-STATE SIMULATION

3.1 Model Assumptions

Our results depend on the following assumptions: For fixed i, Y_{ij}, $j = 1, 2, \ldots, n_i$, is a stationary AR(p_i) process, and Y_{ij} is independent of $Y_{\ell m}$ for $i \neq \ell$. Furthermore, $\sigma_i^2 = \sigma^2$, for $i = 1, 2, \ldots, k$. In other words, initial-condition bias has been mitigated, the dependence in each output process has an autoregressive structure, the systems are simulated independently, and the residual variance is common across systems. The implications of these assumptions are explored in Section 5.

3.2 The MCB Procedure

When $\min_i\{n_i\}$ is large and the assumptions in Section 3.1 are approximately satisfied, Theorem 2.4 justifies the approximation:

$$\begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \vdots \\ \bar{Y}_k \end{bmatrix} \sim N_k\left(\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{bmatrix}, \sigma^2 \begin{bmatrix} \delta_1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_k \end{bmatrix}\right),$$

where $\delta_i = 1/[n_i(1 - \sum_{m=1}^{p_i} \phi_{im})^2]$. This is structure (1). Thus, an AR model provides a natural interpretation for σ^2 and δ_i, which was our primary reason for selecting an AR representation over the more commonly used method of batch means.

For system i, let $\hat{p}_i$ be Rissanen's order estimator, and let $\hat{\phi}_{im} = \hat{\phi}_{im}(n_i, \hat{p}_i)$, for $m = 1, 2, \ldots, \hat{p}_i$, be the combined CLSE and PLS coefficient estimators. An asymptotically valid MCB procedure is formed as follows: Let

$$\hat{\delta}_i = \left[n_i \left(1 - \sum_{m=1}^{\hat{p}_i} \hat{\phi}_{im} \right)^2 \right]^{-1},$$

$$U_i(\mathbf{n}) = \left(\min_{\ell \neq i} \left\{ \overline{Y}_i - \overline{Y}_\ell + \hat{d}^i_{1-\alpha,f} \hat{\sigma} \sqrt{\hat{\delta}_i + \hat{\delta}_\ell} \right\} \right)^+,$$

$$\mathscr{C} = \left\{ i: \min_{\ell \neq i} \left\{ \overline{Y}_i - \overline{Y}_\ell + \hat{d}^i_{1-\alpha,f} \hat{\sigma} \sqrt{\hat{\delta}_i + \hat{\delta}_\ell} \right\} > 0 \right\},$$

$$L_i(\mathbf{n}) = \begin{cases} 0, & \text{if } \mathscr{C} = \{i\}, \\ -\left(\min_{\ell \in \mathscr{C}, \ell \neq i} \left\{ \overline{Y}_i - \overline{Y}_\ell - \hat{d}^\ell_{1-\alpha,f} \hat{\sigma} \sqrt{\hat{\delta}_\ell + \hat{\delta}_i} \right\} \right)^-, & \text{otherwise,} \end{cases}$$

and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^k f_i \hat{\sigma}_i^2}{\sum_{i=1}^k f_i},$$

where $\hat{\sigma}_i^2$ is the estimated residual variance of system i and where f_i is the degrees of freedom associated with $\hat{\sigma}_i^2$ (discussed below). The critical value, $\hat{d}^i_{1-\alpha,f}$, is the $1 - \alpha$ quantile of the maximum of a $(k-1)$-dimensional multivariate t random variable with correlation matrix (3), *where we substitute* $\hat{\delta}_i$ *for* δ_i *in* (4). We set the degrees of freedom for the multivariate t to be $f = \sum_{i=1}^k f_i$.

Our main results follow. For simplicity we present only the balanced case, $n_1 = n_2 = \cdots = n_k = n$:

LEMMA 3.1 *If the assumptions in Section* 3.1 *hold, then as* $n \to \infty$ *the critical value* $\hat{d}^i_{1-\alpha,f} \xrightarrow{\mathscr{D}} d^i_{1-\alpha,\infty}$, *the* $1 - \alpha$ *quantile of the maximum of a* $(k-1)$*-dimensional standard multivariate normal random variable with mean zero and correlation matrix* (3), *where* $\delta_i = 1/(1 - \sum_{m=1}^{p_i} \phi_{im})^2$.

PROOF. The critical value, $\hat{d}^i_{1-\alpha,f}$, is the $1 - \alpha$ quantile of the maximum of a $(k-1)$-dimensional multivariate t random variable with correlation matrix

$$\hat{\mathbf{R}}_i = \begin{bmatrix} 1 & \hat{\lambda}_1^{(i)} \hat{\lambda}_2^{(i)} & \cdots & \hat{\lambda}_1^{(i)} \hat{\lambda}_{k-1}^{(i)} \\ \hat{\lambda}_2^{(i)} \hat{\lambda}_1^{(i)} & 1 & \cdots & \hat{\lambda}_2^{(i)} \hat{\lambda}_{k-1}^{(i)} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\lambda}_{k-1}^{(i)} \hat{\lambda}_1^{(i)} & \hat{\lambda}_{k-1}^{(i)} \hat{\lambda}_2^{(i)} & \cdots & 1 \end{bmatrix},$$

where

$$1/\hat{\lambda}_{\ell}^{(i)} = \begin{cases} \sqrt{1 + \hat{\delta}_{\ell}/\hat{\delta}_i} & \text{if } \ell < i, \\ \sqrt{1 + \hat{\delta}_{\ell+1}/\hat{\delta}_i} & \text{if } \ell \geq i. \end{cases}$$

As a consequence of Theorem 2.4, $\hat{\mathbf{R}}_i$ converges in probability to $\mathbf{R}_i$. And, as $n \to \infty$, the degrees of freedom $f \to \infty$, so the multivariate t distribution converges to the standard multivariate normal distribution. Thus, in the limit, $\hat{d}^i_{1-\alpha,f}$ is the $1 - \alpha$ quantile of the maximum of a $(k-1)$-dimensional standard multivariate normal random variable with correlation matrix $\mathbf{R}_i$. □

The proof of Lemma 3.1 does not depend on the choice of degrees of freedom, f_i, provided f_i goes to infinity as n_i does. However, the choice of f_i does affect the small-sample properties of the MCB procedure. Yuan and Nelson [14] reviewed three alternatives. Of these alternatives, Fishman's approximation is not appropriate here, since it does not provide a basis for pooling the residual variances; the other two are applicable. Based on empirical evidence in [14], we suggest using the full sample size, $f_i = n_i$, when n_i is very large, and using the equivalent sample size, $f_i = \max\{1, \lfloor n(1 - \sum_{m=1}^{\hat{p}_i} \hat{\phi}_{im})^2 \rfloor\}$, when the n_i's are small or when the output processes are strongly positively correlated.

THEOREM 3.1 *If the model assumptions in Section* 3.1 *hold, then as* $n \to \infty$

$$\Pr\left\{\theta_i - \max_{\ell \neq i} \theta_\ell \in [L_i(\mathbf{n}), U_i(\mathbf{n})], \forall i\right\} \to 1 - \alpha.$$

PROOF. By analogy to Hsu's [10] proof of Theorem 2.1, we only need to show that

$$\lim_{n \to \infty} \Pr\left\{\bar{Y}_i - \theta_i \geq \bar{Y}_\ell - \theta_\ell - \hat{d}^i_{1-\alpha,f}\hat{\sigma}\sqrt{\hat{\delta}_i + \hat{\delta}_\ell}, \forall \ell \neq i\right\} = 1 - \alpha,$$

for $i = 1, 2, \ldots, k$. As a direct consequence of Theorem 2.4,

$$\begin{bmatrix} \sqrt{n}\left(\bar{Y}_1 - \theta_1\right) \\ \sqrt{n}\left(\bar{Y}_2 - \theta_2\right) \\ \vdots \\ \sqrt{n}\left(\bar{Y}_k - \theta_k\right) \end{bmatrix} \xrightarrow{\mathcal{D}} \mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix such that $\boldsymbol{\Sigma}(\ell, \ell) = \sigma^2/(1 - \sum_{m=1}^{p_\ell} \phi_{\ell m})^2$.

Fix i and let

$$Z_{\ell,i} = \frac{\left(\bar{Y}_\ell - \theta_\ell\right) - \left(\bar{Y}_i - \theta_i\right)}{\hat{\sigma}\sqrt{\hat{\delta}_\ell + \hat{\delta}_i}}.$$

Define a $(k-1) \times k$ transformation matrix $\mathbf{A}_i$ such that $\mathbf{A}_i(\ell, i) = -1$, for all ℓ; $\mathbf{A}_i(\ell, \ell) = 1$, for $\ell = 1, 2, \ldots, i-1$; $\mathbf{A}_i(\ell, \ell+1) = 1$, for $\ell = i, i+$

$1, \ldots, k-1$; and all remaining elements of $\mathbf{A}_i$ are 0. Also, define a $(k-1) \times (k-1)$ diagonal matrix $\hat{\mathbf{V}}_i$ such that $\hat{\mathbf{V}}_i(\ell, \ell) = (n\hat{\sigma}^2(\hat{\delta}_i + \hat{\delta}_\ell))^{-1/2}$, for $\ell < i$, and $\hat{\mathbf{V}}_i(\ell, \ell) = (n\hat{\sigma}^2(\hat{\delta}_i + \hat{\delta}_{\ell+1}))^{-1/2}$, for $\ell \geq i$. Then

$$\begin{bmatrix} Z_{1,i} \\ Z_{2,i} \\ \vdots \\ Z_{i-1,i} \\ Z_{i+1,i} \\ \vdots \\ Z_{k,i} \end{bmatrix} = \hat{\mathbf{V}}_i \mathbf{A}_i \begin{bmatrix} \sqrt{n}\left(\overline{Y}_1 - \theta_1\right) \\ \sqrt{n}\left(\overline{Y}_2 - \theta_2\right) \\ \vdots \\ \sqrt{n}\left(\overline{Y}_k - \theta_k\right) \end{bmatrix}.$$

Since $\hat{\mathbf{V}}_i \xrightarrow{\mathcal{D}} \mathbf{V}_i$, where $\mathbf{V}_i(\ell, \ell) = (n\sigma^2(\delta_i + \delta_\ell))^{-1/2}$, for $\ell < i$, and $\mathbf{V}_i(\ell, \ell) = (n\sigma^2(\delta_i + \delta_{\ell+1}))^{-1/2}$, for $\ell \geq i$, we have

$$\begin{bmatrix} Z_{1,i} \\ Z_{2,i} \\ \vdots \\ Z_{i-1,i} \\ Z_{i+1,i} \\ \vdots \\ Z_{k,i} \end{bmatrix} \xrightarrow{\mathcal{D}} \mathrm{N}_{k-1}(\mathbf{0}, \mathbf{V}_i \mathbf{A}_i \boldsymbol{\Sigma} \mathbf{A}_i' \mathbf{V}_i').$$

But $\mathbf{V}_i \mathbf{A}_i \boldsymbol{\Sigma} \mathbf{A}_i' \mathbf{V}_i' = \mathbf{R}_i$. Therefore, using the result of Lemma 3.1,

$$\begin{aligned} &\lim_{n\to\infty} \Pr\left\{\overline{Y}_i - \theta_i \geq \overline{Y}_\ell - \theta_\ell - \hat{d}^i_{1-\alpha,f}\hat{\sigma}\sqrt{\hat{\delta}_i + \hat{\delta}_\ell}, \forall \ell \neq i\right\} \\ &\quad = \lim_{n\to\infty} \Pr\left\{Z_{\ell,i} \leq \hat{d}^i_{1-\alpha,f}, \forall \ell \neq i\right\} \\ &\quad = \lim_{n\to\infty} \Pr\left\{\max_{\ell \neq i}\{Z_{\ell,i}\} \leq \hat{d}^i_{1-\alpha,f}\right\} \\ &\quad = \Pr\{X \leq d^i_{1-\alpha,\infty}\} \\ &\quad = 1 - \alpha, \end{aligned} \tag{7}$$

where X is the maximum of a $(k-1)$-dimensional standard multivariate normal random variable with correlation matrix $\mathbf{R}_i$. Given (7), we can directly follow the steps in [10] to show that $\{[L_1(\mathbf{n}), U_1(\mathbf{n})], [L_2(\mathbf{n}), U_2(\mathbf{n})], \ldots, [L_k(\mathbf{n}), U_k(\mathbf{n})]\}$ is asymptotically a set of $(1-\alpha)100$ percent confidence intervals for $\theta_i - \max_{\ell \neq i} \theta_\ell$, $i = 1, 2, \ldots, k$. □

Theorem 3.1 can be extended to unbalanced cases ($n_i \neq n_\ell$), provided $\lim_{n_\ell, n_i \to \infty} n_\ell / n_i$ is finite for all $\ell \neq i$. To implement the procedure, we need to set aside storage space for the k output processes. We recommend batching the output processes first, which reduces the storage space required and tends to improve the AR approximation; see test set 4 in the next section. A routine is needed to evaluate the AR order and coefficients, the residual variance and associated degrees of freedom, and the sample mean. These

results are stored and passed to another routine to compute MCB confidence intervals. We wrote our own MCB routines, but several commercial statistical-analysis packages also contain them, including JMP version 2 and Minitab release 8. The procedure can be readily automated, since no user judgment is required to select the AR order.

The PLS criterion could be computationally expensive to use, since a matrix inversion is required for each candidate order and each observation. In [14] we presented a procedure for computing $\hat{p}_i$ that efficiently updates a single matrix inversion for each candidate order, rather than performing the inversion; this procedure also yields the AR coefficient estimators as a by-product.

4. EMPIRICAL STUDY

For a set of test cases, we estimated the probability that our MCB intervals are correct (cover all of the parameters) and that they are both correct and conclusive (cover all of the parameters and identify differences between systems' performance) for nominal 90 percent MCB intervals. The probability of being "correct" should be 0.90, and this is the probability that we expect the procedure to control. The probability of being "correct and conclusive" is the probability of identifying the best system design as well as covering all of the MCB differences; it is a function of the sample size and was included in the study to obtain a sense of the sharpness of the procedure for identifying the best.

The estimates were obtained by replicating each experiment 100 times; thus, the first digit in each estimate is accurate, but the second less certain. The simulations were initialized using the steady-state distribution, when it was known; otherwise, an appropriate number of initial outputs were deleted. Three levels of sample size, n, were used to compare the small-sample and large-sample behavior of the procedure. The set of possible AR orders was $\mathcal{O} = \{1, 2, \ldots, 8\}$. In addition, we used both the full sample size and the equivalent sample size to obtain the degrees of freedom, f.

4.1 Test Sets

Four sets of processes were selected for experiments. In all cases the objective was to identify the system with the largest mean response from among $k = 5$ competing systems.

(1) *Test Set* 1. In this set we used AR processes. They are

$$
\begin{aligned}
Y_{1,j} - \theta_1 &= 0.3(Y_{1,j-1} - \theta_1) + \epsilon_{1,j},\\
Y_{2,j} - \theta_2 &= 0.8(Y_{2,j-1} - \theta_2) + \epsilon_{2,j},\\
Y_{3,j} - \theta_3 &= 0.5(Y_{3,j-1} - \theta_3) + 0.25(Y_{3,j-2} - \theta_3) + \epsilon_{3,j},\\
Y_{4,j} - \theta_4 &= 0.3(Y_{4,j-1} - \theta_4) + 0.2(Y_{4,j-2} - \theta_4) + 0.1(Y_{4,j-3} - \theta_4) + \epsilon_{4,j},\\
Y_{5,j} - \theta_5 &= 0.4(Y_{5,j-1} - \theta_5) + 0.2(Y_{5,j-2} - \theta_5) + 0.1(Y_{5,j-3} - \theta_5)\\
&\quad + 0.05(Y_{5,j-4} - \theta_5) + \epsilon_{5,j},
\end{aligned}
$$

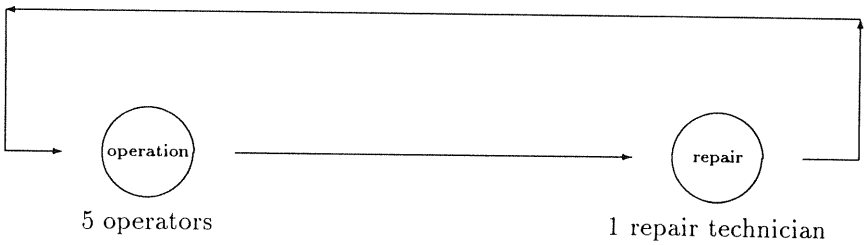

Fig. 2. Production system for test set 4.

where $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (5.0, 5.25, 5.50, 5.75, 6.00)$ and where the $\epsilon_{i,j}$'s are i.i.d. N(0, 1) random variables, $\forall i, j$.

(2) *Test Set* 2. In this set we used models with nonnormal marginal distributions to observe the effect of deviation from normality. The test models are identical to set 1 except that the $(\epsilon_{i,j} + 1)$'s are i.i.d. exponential random variables with mean 1, $\forall i, j$. The revised models are still covariance stationary.

(3) *Test Set* 3. In this set we used ARMA processes to observe the effect of correlation structures different from AR processes. They are

$$Y_{1,j} - \theta_1 = 0.7(Y_{1,j-1} - \theta_1) + \epsilon_{1,j} + 0.5\epsilon_{1,j-1},$$

$$Y_{2,j} - \theta_2 = 0.3(Y_{2,j-1} - \theta_2) + \epsilon_{2,j} + 0.2\epsilon_{2,j-1},$$

$$Y_{3,j} - \theta_3 = 0.5(Y_{3,j-1} - \theta_3) + 0.25(Y_{3,j-2} - \theta_3) + \epsilon_{3,j} + 0.3\epsilon_{3,j-1},$$

$$Y_{4,j} - \theta_4 = 0.4(Y_{4,j-1} - \theta_4) + 0.2(Y_{4,j-2} - \theta_4) + \epsilon_{4,j} + 0.25\epsilon_{4,j-1} + 0.1\epsilon_{4,j-2},$$

$$Y_{5,j} - \theta_5 = 0.35(Y_{5,j-1} - \theta_5) + 0.25(Y_{5,j-2} - \theta_5) + 0.15(Y_{5,j-3} - \theta_5) + \epsilon_{5,j} + 0.2\epsilon_{5,j-1},$$

where $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (5.3, 5.0, 5.6, 5.9, 6.2)$, and the $\epsilon_{i,j}$'s are i.i.d N(0, 1) random variables, $\forall i, j$.

(4) *Test Set* 4. In this set we used the production system shown in Figure 2. The system consists of five identical machines, n spares, five machine operators, and one repair technician. The failure times and repair times are exponentially distributed random variables. There are five alternatives that trade off spare machines against reduced mean repair time in order to maximize the long-run expected number of machines in operation. The five alternatives are listed in Table I. The output Y_{ij} is the average number of operating machines over the jth batch of 40 time units for alternative i; that is, we used batch means of the original output process as the basic output data.

Table I. Investment Alternatives for the Production System in Test Set 4

Alternative	Spare machines	Mean repair time
1	4	3.25
2	3	3.00
3	2	2.50
4	1	2.00
5	0	1.75

Table II. Number of MCB Intervals Out of 100 that Were Correct or Correct and Conclusive

	Sample Size					
	500		2500		5000	
Degrees of freedom	Full sample	Equivalent sample	Full sample	Equivalent sample	Full sample	Equivalent sample
Test set 1						
Correct	91	91	87	88	95	95
Correct and conclusive	27	27	71	70	95	95
Test set 2						
Correct	90	90	88	88	89	89
Correct and conclusive	22	22	70	69	89	89
Test set 3						
Correct	75	75	94	94	96	96
Correct and conclusive	17	17	79	79	95	95
Fail to compute $d_{1-\alpha}$	10	10	0	0	0	0
Test set 4						
Correct	89	87	93	91	88	87
Correct and conclusive	32	33	71	71	88	87

4.2 Experiment Results

In test set 1 (see Table II), we find that the probability of correct inference is close to the nominal level at all three sample sizes. The probability of correct and conclusive inference is reasonable only in the moderate- or large-sample case. Almost no difference is observed between the two approximations for degrees of freedom. When the underlying assumptions are satisfied, the procedure seems able to meet the nominal coverage. A moderate sample size may be required for the procedure to reach a conclusive decision.

In test set 2, we find that the probability of correct inference is still close to the nominal level. But the probability of correct and conclusive inference is a bit lower as compared to test set 1. Thus, the deviation from the normal residual assumption did not have a significant effect.

In test set 3, for small n, we encountered a few cases where we failed to compute MCB critical values because the numerical integration and search procedure did not converge. Also, the probability of correct inference is much lower than 0.90 when the sample size is small. This suggests the deviation from AR-correlation structure may have adverse effects (an ARMA model is

an infinite-order AR model). The procedure works well in the moderate- or large-sample case.

In test set 4, the overall performance is satisfactory. Similar trends are observed: The probability of correct inference is usually under control. When the sample size is small, the probability of correct and conclusive inference is low, but it improves with increased sampling. Again, there seems to be no difference between the MCB intervals computed using the two approximations for degrees freedom.

5. CONCLUSIONS AND DISCUSSION

The asymptotic validity of our procedure (Theorem 3.1) depends on the assumption that the underlying output processes are AR(p), which is never precisely true in practice. However, provided that we allow the set of candidate orders, $\mathscr{O}$, to be large enough, an AR(p) model can provide a good approximation to the output processes, as it did in the empirical study.

A more critical assumption is that the residual variances are common across systems ($\sigma_i^2 = \sigma^2$, for $i = 1, 2, \ldots, k$). This should be empirically verified in practice. Notice, however, that we do not assume that the response variances, $\mathrm{Var}[Y_{ij}]$, or the variances of the sample means, $\mathrm{Var}[\overline{Y}_{ij}]$, are equal. Instead, we assume that the remaining variance, after removing the effect of the AR-dependence structure, is the same across systems. Batching can be used to improve this approximation when it is violated.

We also assume that the k systems are simulated independently; in practice, this means that different random number streams are used to drive the simulation of each system. This assumption rules out the use of common random numbers (CRN), which is known to reduce the variance of estimators of differences (such as $\overline{Y}_i - \max_{\ell \neq i} \overline{Y}_\ell$). Application of CRN in this setting is still an open problem.

Some tentative conclusions can be drawn from the empirical study:

(1) In most of the experiments, the MCB procedure was able to control the probability of correct inference. This supports the validity of the procedure in small samples.

(2) The MCB procedure seemed to be robust to the assumption of normal residuals or AR correlation structure when the sample size was at least moderate.

(3) The two approximations for degrees of freedom lead to similar performance. We conjecture that the approximate degrees of freedom for the pooled-residual-variance estimator in either case is so large that it has little effect on the MCB critical values.

In closing, we mention that, even though this paper focused on MCB inference, the method we propose extends directly to multiple comparisons with a control system.

REFERENCES

1. AKAIKE, H. Maximum likelihood identification of Gaussian auto-regressive moving average process. *Biometrika 60*, 2 (Aug. 1973), 255–266.

2. BOX, G. E. P., AND JENKINS, G. M. *Time Series Analysis*. Rev. ed. Holden-Day, San Francisco, Calif., 1976.
3. FISHMAN, G. S. *Principles of Discrete Event Simulation*. Wiley, New York, 1973.
4. GRAY, H. L., KELLY, G. D., AND MCINTIRE, D. D. A new approach to ARMA modelling. *Commun. Stat. B7* (1978), 1–77.
5. HANNAN, E. J. *Multiple Time Series*. Wiley, New York, 1970.
6. HANNAN, E. J., MCDOUGALL, A. J., AND POSKITT, D. S. Recursive estimation of autoregressions. *J. Roy. Stat. Soc. B51*, 2 (1989), 217–233.
7. HEMERLY, E. M., AND DAVIS M. H. A. Strong consistency of the PLS criterion for order determination of autoregressive processes. *Ann. Stat. 17*, 2 (1989), 941–946.
8. HOCHBERG, Y., AND TAMHANE, A. C. *Multiple Comparison Procedures*. Wiley, New York, 1987.
9. HSU, J. C. Constrained simultaneous confidence intervals for multiple comparisons with the best. *Ann. Stat. 12*, 3 (1984), 1136–1144.
10. HSU, J. C. Ranking and selection and multiple comparisons with the best. In *Design of Experiments: Ranking and Selection*, T. J. Santner and A. C. Tamhane, Eds. Marcel Dekker, New York, 1984, chap. 3.
11. HSU, J. C., AND NELSON, B. L. Optimization over a finite number of system designs with one-stage sampling and multiple comparisons with the best. In *Proceedings of the 1988 Winter Simulation Conference* (San Diego, Calif., Dec. 12–14). ACM/IEEE, New York, 1988, 451–457.
12. NELSON, B. L., AND HSU, J. C. Control-variate models of common random numbers for multiple comparisons with the best. *Manage. Sci.* To be published.
13. RISSANEN, J. Order estimation by accumulated prediction errors. In *Essays in Time Series and Allied Processes*. J. Gani and M. B. Priestly, Eds. *J. Appl. Prob. 23A* (1986), 55–61.
14. YUAN, M., AND NELSON, B. L. Autoregressive output analysis methods revisited. Working Paper Ser. 1991-007, Dept. of Industrial and Systems Engineering, Ohio State Univ., Columbus, 1991.