Autoregressive-output-analysis methods revisited

Mingjian Yuan

Department of Industrial Management, National Yunlin Institute of Technology, Yunlin, Taiwan, Republic of China

Barry L. Nelson

Department of Industrial and Systems Engineering, The Ohio State University, Columbus, OH 43210, USA

We revisit and update the autoregressive-ouput-analysis method for constructing a confidence interval for the steady-state mean of a simulated process by using Rissanen's predictive least-squares criterion to estimate the autoregressive order of the process. This order estimator is strongly consistent when the output is autoregressive. The order estimator is combined with the standard autoregressive-output-analysis method to form a confidence-interval procedure. Alternatives for estimating the degrees of freedom for the procedure are investigated. The main result is an asymptotically valid confidence-interval procedure that, empirically, has good small-sample properties.

Keywords: Autoregressive process, confidence interval, output analysis, simulation, statistics, time series.

1. Introduction

A standard experiment design for steady-state simulation is to allocate the entire computing budget to a single long run. This design reduces data waste and the potential mistakes from initial-transient deletion. However, statistical analysis of the output process is a longstanding problem. We present a parametric method for deriving a confidence interval for the mean of a stationary stochastic process. The method models the output process using an autoregressive representation, and the statistical analysis exploits the well-known properties of autoregressive models. This is an old idea that we update substantially. In particular, we use Rissanen's predictive least-squares criterion to estimate the autoregressive order of the output process. This order estimator is strongly consistent when the output process is autoregressive, and we combine it with the standard autoregressive-output-analysis method to form an asymptotically valid confidence-interval procedure. Alternatives for estimating the degrees of freedom for the procedure are also investigated.

In the next section we introduce and review autoregressive modeling. In section 3 we present the underlying methodology and our main result: an asymptotically valid confidence-interval procedure. We empirically examine the robustness and small-sample properties of this procedure in section 4; conclusions and recommendations are given in section 5.

2. Background

Let Y_j denote the jth output from a single replication. An autoregressive order p model, denoted AR(p), approximates the dependence in $\{Y_j;\ j = 1, 2, \ldots\}$ by the linear autoregression

$$Y_j - \theta = \sum_{i=1}^{p} \phi_i(Y_{j-i} - \theta) + \varepsilon_j, \tag{1}$$

where the ϕ_i's are the autoregressive coefficients, θ is the unknown process mean, and the ε_j's are independent and identically distributed (i.i.d.) residuals with zero mean and finite variance σ^2. There are three considerations that motivate an AR-modeling approach:

1. An AR model is a reasonable approximation for general output processes. Given that the autocovariance matrix of an output process is invertible, we can approximate the output process by an AR model that matches the autocovariance to as many lags as desired.
2. Theoretical development of AR modeling is complete, relative to autoregressive-moving average (ARMA) modeling, in the sense that a strongly consistent order estimator has been developed for AR models but not for ARMA models. This property is critical to developing an asymptotically valid confidence-interval procedure; no such procedure existed using AR or ARMA modeling prior to this work.
3. A simulation experiment seldom ends with analyzing a single system. One often needs to compare alternatives and select the best. An AR method is appropriate in conjunction with multiple-comparison procedures because AR models provide a natural interpretation for the assumptions behind multiple-comparison procedures, while other well-known output-analysis methods (e.g. batch means) do not. Yuan and Nelson [21] derived the first asymptotically valid multiple-comparison procedure for steady-state simulation based upon the theoretical foundation in this paper.

Fishman [7] originated the application of AR modeling in simulation analysis. Following his initial work, research moved toward ARMA models as an alternative because of their parsimonious representation. This paper revisits Fisherman's procedure. The goal is to incorporate newly developed methods for AR order identification and estimation of degrees of freedom to derive a provably valid procedure.

3. Methodology

This section presents an AR-confidence-interval procedure for steady-state simulation. Figure 1 is a flowchart of the proposed procedure. In the subsections that follow we fill in the details for each step in the chart. To keep the presentation concise, we begin with AR coefficient estimation given that the AR order is known, and then describe how the order is determined. The proofs of all lemmas and theorems that are not referenced can be found in the appendix.

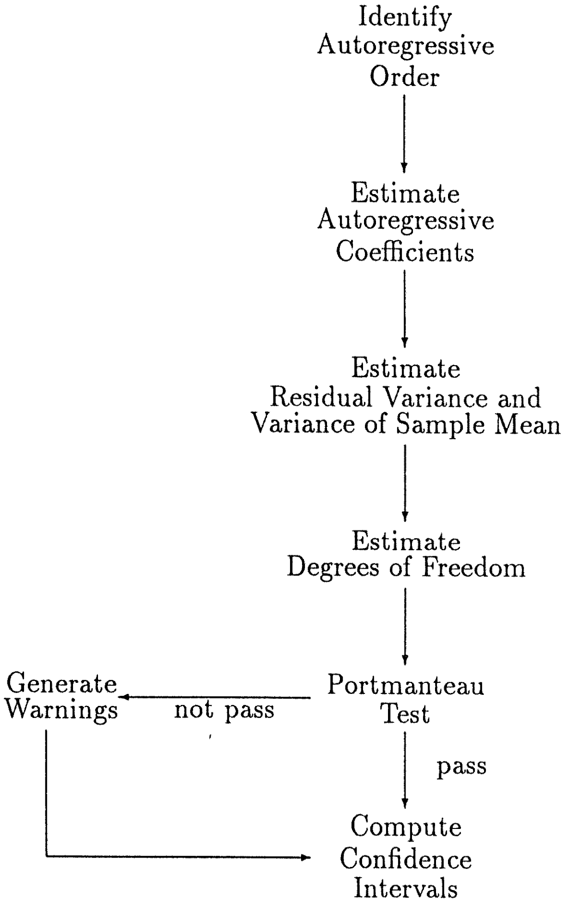
Identify Autoregressive Order
Estimate Autoregressive Coefficients
Estimate Residual Variance and Variance of Sample Mean
Estimate Degrees of Freedom
Portmanteau Test
not pass
Generate Warnings
pass
Compute Confidence Intervals

Figure 1. AR-confidence-interval procedure.

Throughout the remainder of this paper we assume the output process $\{Y_j; j = 1, 2, \ldots\}$ is a stationary AR(p) process represented by model (1). The ϕ_i's in the model are assumed to satisfy the conditions that insure $\{Y_j\}$ to be covariance stationary (i.e. the p roots of the characteristic equation $1 - \sum_{i=1}^{p} \phi_i x^i = 0$ all exceed unity in modulus; see [5, p. 70]). A useful reparameterization of the model is

$$Y_j = \phi_0 + \sum_{i=1}^{p} \phi_i Y_{j-1} + \varepsilon_j, \tag{2}$$

where $\phi_0 = \theta(1 - \sum_{i=1}^{p} \phi_i)$.

3.1. AR COEFFICIENT ESTIMATION

When the order of an AR process p is known, there are a number of coefficient estimators available in the literature. They include the Yule–Walker estimator (YWE), the maximum-likelihood estimator (MLE) and the conditional-least-squares estimator (CLSE); the latter is a least-squares estimator conditional on the initial p observations of the Y_j's (see e.g. [5]). We propose using CLSE for the following reasons:

1. Given p, both CLSE and YWE yield consistent estimators, but CLSE usually has a smaller first-order asymptotic bias [18]; i.e. CLSE converges to the true parameters faster.
2. The computation of CLSE is straightforward, and CLSE is a by-product of the order identification procedure we use. However, directly computing the MLEs can be difficult and one typically needs to use some approximation [3].

In the following we summarize the computation and properties of CLSE.

3.1.1. *Computation*

Let the CLSE be denoted by $\hat{\phi}(n, p) = [\hat{\phi}_0(n, p), \hat{\phi}_1(n, p), \ldots, \hat{\phi}_p(n, p)]'$, where n is the sample size. Define the data and design matrices

$$\mathbf{Y}(n, p) = [Y_{p+1}, Y_{p+2}, \ldots, Y_n]',$$

$$\mathbf{X}(n, p) = \begin{bmatrix} 1 & Y_p & Y_{p-1} & \cdots & Y_1 \\ 1 & Y_{p+1} & Y_p & \cdots & Y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Y_{n-1} & Y_{n-2} & \cdots & Y_{n-p} \end{bmatrix}.$$

Then

$$\hat{\phi}(n, p) = [\mathbf{X}'(n, p)\,\mathbf{X}(n, p)]^{-1}\mathbf{X}'(n, p)\,\mathbf{Y}(n, p)$$

and

$$\theta'(n, p) = \hat{\phi}_0(n, p) \Big/ \left[1 - \sum_{i=1}^{p} \hat{\phi}_i(n, p)\right],$$

where $\theta'(n, p)$ is a point estimator of the process mean θ.

3.1.2. Properties

Let $\xrightarrow{\mathscr{P}}$ denote convergence in probability. The following lemmas show that the CLSE coefficient estimators are consistent when the order p is known.

LEMMA 3.1 [5, p. 130]

If the model assumptions in section 3 hold, and the order p is known, then $\hat{\theta}(n, p) - \bar{Y} \xrightarrow{\mathscr{P}} 0$ (implying that $\hat{\theta}(n, p) \xrightarrow{\mathscr{P}} \theta$).

LEMMA 3.2 [20, pp. 147–149]

If the model assumptions in section 3 hold, and the order p is known, then $\hat{\phi}_i(n, p) \xrightarrow{\mathscr{P}} \phi_i$, for $i = 1, 2, \ldots, p$.

3.2. AR ORDER IDENTIFICATION

Determining the order of an AR process has been a longstanding problem. There are various perspectives in the literature. One perspective treats order identification as a hypothesis-testing problem. An F test was developed by Hannan [10]. Fishman [7] adopted this test in his AR procedure. Box and Jenkins [3] treat order identification as an iterative procedure: Governed by a principle of parsimony, a candidate model is selected by investigating correlograms. Then diagnostic tests are performed. The candidate models is iteratively modified until it passes all tests. Another perspective is to use a single numerical index to measure the goodness of a model. One such index is Akaike's [1] information criterion. Gray et al. [9] developed two heuristic "D" statistics. These statistics were adopted by Schriber and Andrews [17] and Chun [4]. None of these approaches has been shown to provide an asymptotically consistent order estimator.

More recently, Rissanen [16] proposed the predictive-least-squares (PLS) criterion. PLS selects an order based on the predictive ability of a model. We adopt this criterion and develop an efficient implementation of it.

For candidate order h, let

$$\mathrm{PLS}_h = \frac{\sum_{i=2h+2}^{n} e_i^2(h)}{n - 2h - 1}, \tag{3}$$

where $e_i^2(h)$ is the "honest prediction error" for Y_i asuming the order to be h. That is, $e_i(h) = Y_i - \hat{Y}_i(h)$, where $\hat{Y}_i(h)$ is the predicted value of Y_i from the AR(h) model estimated by CLSE, but using only observations $Y_1, Y_2, \ldots, Y_{i-1}$ (therefore it is "honest" in the sense that only observations prior to Y_i enter into the predicted value of Y_i). Following the notation defined in subsection 3.1.1,

$$\hat{Y}_i(h) = \{[\mathbf{X}'(i-1,h)\mathbf{X}(i-1,h)]^{-1}\mathbf{X}'(i-1,h)\mathbf{Y}(i-1,h)\}' \begin{bmatrix} 1 \\ Y_{i-1} \\ Y_{i-2} \\ \vdots \\ Y_{i-h} \end{bmatrix}.$$

The PLS criterion selects the order $\hat{p}$ from a set $\mathcal{O}$ of possible orders such that $\text{PLS}_{\hat{p}} = \min_{h \in \mathcal{O}} \text{PLS}_h$. The set of candidate orders $\mathcal{O}$ is assumed to contain the true order p. Notice that $e_i(h)$ can be computed only for $i \geq 2h + 2$.

Several papers discuss the properties of PLS. Rissanen [16] showed that $\hat{p}$ is consistent in the case of Gaussian residuals. Wax [19] showed that $\hat{p}$ is consistent without the Gaussian assumption. Hannan et al. [11] showed $\hat{p}$ to be strongly consistent. Hemerly and Davis [12] derived the same conclusion. This convergence property is critical to developing an asymptotically valid confidence-interval procedure.

A naive implementation of the PLS criterion is computationally expensive to use because a matrix inversion is required for each candidate order, and for each observation. We develop an efficient algorithm in the appendix.

Let $\hat{\phi}(n, \hat{p})$ be the CLSE coefficient estimator in conjunction with Rissanen's order estimator $\hat{p}$; i.e. $\hat{\phi}(n, \hat{p}) = \hat{\phi}(n, p)$ when $\hat{p} = p$. Using lemmas 3.1–3.3, we can show that $\hat{\phi}(n, \hat{p})$ is consistent for ϕ.

LEMMA 3.3

If the model assumptions in section 3 hold, then for each $\varepsilon > 0$,

$$\lim_{n \to \infty} \Pr\{|\hat{\phi}_i(n,p) - \phi_i| < \varepsilon \text{ and } \hat{p} = p\} = 1, \quad \text{for } i = 1, 2, \ldots, p.$$

THEOREM 3.1

If the model assumptions in section 3 hold, then $\hat{\phi}(n, \hat{p}) \xrightarrow{\mathcal{P}} \phi$ as $n \to \infty$.

Theorem 3.1 states that the CLSE coefficient estimator combined with the PLS order estimator is consistent. A corollary (that we will need later) follows directly:

COROLLARY 3.1

If the model assumptions in section 3 hold, then $\sum_{i=1}^{\hat{p}} \hat{\phi}_i(n, \hat{p}) \xrightarrow{\mathcal{P}} \sum_{i=1}^{p} \phi_i$ as $n \to \infty$.

3.3. VARIANCE ESTIMATION

The confidence interval we propose is of the form

$$\bar{Y} \pm t_{1-\alpha/2,f} \sqrt{\widehat{\text{Var}}[\bar{Y}]},$$

where $t_{1-\alpha/2,f}$ is the $1 - \alpha/2$ quantile of a t random variable with f degrees of freedom. Given that $\{Y_j, j = 1,2, \ldots, n\}$ is a stationary AR(p) process, it is known that for large sample size n,

$$\text{Var}[\bar{Y}] \approx \frac{\sigma^2}{n(1 - \sum_{i=1}^{p} \phi_i)^2} \tag{4}$$

[8]. Therefore, a natural estimator of $\text{Var}[\bar{Y}]$ is $\widehat{\text{Var}}[\bar{Y}]$ that is obtained by substituting estimates for the parameters in (4). If we approximate the distribution of $\widehat{\text{Var}}[\bar{Y}]$ as a constant times a χ^2 random variable, and if we are given an estimate of the associated degrees of freedom f, then we can construct an approximate confidence interval for the mean. This section discusses estimation of σ^2 and the degrees of freedom.

3.3.1. Estimation of the residual variance

The residual variance can be estimated by

$$\hat{\sigma}^2 = \frac{\sum_{j=\hat{p}+1}^{n} [Y_j - \hat{\phi}_0(n, \hat{p}) - \sum_{i=1}^{\hat{p}} \hat{\phi}_i(n, \hat{p}) Y_{j-i}]^2}{n - \hat{p}}. \tag{5}$$

Using the results from the previous sections we have the following theorems:

THEOREM 3.2

If the model assumptions in section 3 hold, then $\hat{\sigma}^2 \xrightarrow{\mathcal{P}} \sigma^2$ as $n \to \infty$.

The PLS order estimator and CLSE coefficient estimator can be combined with the residual variance estimator $\hat{\sigma}^2$ to form a variance estimator

$$\widehat{\text{Var}}[\bar{Y}] = \frac{\hat{\sigma}^2}{n(1 - \sum_{i=1}^{\hat{p}} \hat{\phi}_i(n, \hat{p}))^2}.$$

This estimator is consistent, as shown in the next result. Let $\tau^2 = \lim_{n\to\infty} n\mathrm{Var}[\bar{Y}] = \sigma^2/(1 - \sum_{i=1}^{p} \phi_i)^2$.

THEOREM 3.3

If the model assumptions in section 3 hold, then $n\widehat{\mathrm{Var}}[\bar{Y}] \xrightarrow{\mathcal{P}} \tau^2$ as $n \to \infty$.

3.3.2. Degrees of freedom for $\widehat{\mathrm{Var}}[\bar{Y}]$

Determining the degrees of freedom to associate with σ^2 is difficult because the small-sample distribution of σ^2 is not known. Determining the degrees of freedom of $\widehat{\mathrm{Var}}[\bar{Y}]$ is even harder. The degrees of freedom will affect the t critical values that are used in computing confidence intervals. This section introduces approximations for large and small-sample cases. The definition of "large" and "small" depends, unfortunately, on unknown properties of the particular output process at hand. One reason that we examine both approximations in the empirical study (section 4) is to determine whether there is substantial benefit from using the more complicated small-sample approximations.

Large-sample case: Since both the order and the coefficient estimators converge to their true values as the sample size increases, the problem is simplified by treating the estimates as if they are known constants. Then $(n - \hat{p})\hat{\sigma}^2/\sigma^2$ is a χ^2 random variable with $n - \hat{p}$ degrees of freedom, and $(n - \hat{p})\,\widehat{\mathrm{Var}}[\bar{Y}]/\mathrm{Var}[\bar{Y}]$ is approximately a χ^2 random variable with $n - \hat{p}$ degrees of freedom. In this paper, we simply use the full sample size as the large-sample degrees of freedom because n is always much greater than $\hat{p}$.

Small-sample case: When the sample size is moderate or small, taking n or $n - \hat{p}$ as the degrees of freedom for $(n - \hat{p})\,\widehat{\mathrm{Var}}(\bar{Y})/\mathrm{Var}[\bar{Y}]$ is not justified. Fishman [7] suggested the following approximation:

1. Assume that $\widehat{\mathrm{Var}}[\bar{Y}]/\mathrm{Var}[\bar{Y}]$ is a χ^2_f/f random variable.
2. Assumption 1 implies that $2(\mathrm{E}[\widehat{\mathrm{Var}}[\bar{Y}]])^2/\mathrm{Var}[\widehat{\mathrm{Var}}[\bar{Y}]] = f$, so f is approximated by

$$\hat{f} = \frac{2(\widehat{\mathrm{Var}}[\bar{Y}])^2}{\widehat{\mathrm{Var}}[\widehat{\mathrm{Var}}[\bar{Y}]]}.$$

3. The $\widehat{\mathrm{Var}}[\widehat{\mathrm{Var}}(\bar{Y})]$ is a sample estimate based on the following limiting result [7]:

$$\lim_{n\to\infty} n\mathrm{Var}[\widehat{\mathrm{Var}}[\bar{Y}]] = \left[\frac{\sigma^2}{(1 - \sum_{i=1}^{p}\phi_i)^2}\right]^2 \left[2 + 4\left(\frac{p - \sum_{i=1}^{p}\phi_i(p - 2i)}{1 - \sum_{i=1}^{p}\phi_i}\right)\right].$$

After simplification,

$$\hat{f} = \frac{n[1 - \sum_{i=1}^{\hat{p}} \hat{\phi}_i(n, \hat{p})]}{(1 + 2\hat{p})[1 - \sum_{i=1}^{\hat{p}} \hat{\phi}_i(n, \hat{p})] + 4\sum_{i=1}^{\hat{p}} i\hat{\phi}_i(n, \hat{p})}.$$

The quantity $\hat{f}$ is similar to a moment estimator.

A second approximation takes a different view: Consider a stochastic process $\{Z_j; j = 1, 2, \ldots, n'\}$, where $Z_j \sim$ i.i.d. (θ, σ^2) and $n' = \max\{1, n(1 - \sum_{i=1}^{p} \phi_i)^2\}$. Let $\bar{Z} = \sum_{j=1}^{n'} Z_j / n'$. Then $\mathrm{E}[\bar{Z}] = \theta$ and $\mathrm{Var}[\bar{Z}] \approx \mathrm{Var}[\bar{Y}]$. We may regard the information contained in $\{Z_j\}$ to be equivalent to that in $\{Y_j\}$ for purposes of variance estimation. This suggests taking the degrees of freedom of $\widehat{\mathrm{Var}}[\bar{Y}]$ to be the degrees of freedom of $\widehat{\mathrm{Var}}[\bar{Z}]$; i.e. n'. Schriber and Andrews [17] suggested a similar, but slightly more conservative, approximation that was also adopted by Chun [4]. Notice that n' could be larger than or smaller than n, and that n' is exactly n if $\{Y_j\}$ is an i.i.d. process. To obtain n' in practice we substitute estimates for p and the ϕ_i's.

3.3.3. *Confidence-interval procedure*

As discussed in the previous section, there are alternative ways to estimate the degrees of freedom. Assembling different estimators results in different confidence-interval procedures. We will evaluate all three alternatives: Fishman's approximation, the equivalent sample size, and the full sample size.

Theorem 3.4 establishes that our confidence-interval procedure will be asymptotically valid provided that the method for estimating the degrees of freedom forces f to go to infinity as the sample size does.

THEOREM 3.4

If the model assumptions in section 3 hold, then

$$\frac{\bar{Y} - \theta}{\sqrt{\widehat{\mathrm{Var}}[\bar{Y}]}} \Rightarrow \mathrm{N}(0, 1),$$

where N(0, 1) denotes a standard normal random variable and ⇒ denotes convergence in distribution.

3.4. LACK-OF-FIT TEST

Rarely can an output process be perfectly modeled as an AR process. A lack-of-fit test is sometimes recommended as a quality inspection. In the literature, Portmanteau's test is used [14]. The test checks the goodness of a model by computing the Ljung–Box–Pierce statistic, which is given by

$$Q = n(n+2)\sum_{j=1}^{k} \frac{\hat{\gamma}_j^2}{n-k},$$

where $\hat{\gamma}_j$ is the lag-j sample autocorrelation of the estimated residuals from the selected AR(p) model, and k is a value large enough so that $\hat{\gamma}_j$ is negligible for $j > k$. This paper takes k to be the largest integer less than log n. The limiting null distribution of Q is close to a χ^2 distribution with $k-p$ degrees of freedom. We could reject a pth order model if $Q \geq \chi^2_{1-\alpha,k-p}$, the $1-\alpha$ quantile of a χ^2 random variable with $k-p$ degrees of freedom. A more complete discussion of this test can be found in [6]. In the next section we investigate whether it would be a serious error to ignore the test.

4. Experiments

In this section we evaluate the AR-confidence-interval procedures empirically. Five sets of process are selected for experiments. These sets include AR processes with normal residuals, AR processes with (shifted) exponentially distributed residuals, ARMA processes, *M*/*M*/1 queues, and a production system. These sets are chosen to systematically stress the procedures by violating their assumptions, and to test them on systems simulation examples. In addition, three sample sizes are used to compare the small-sample and large-sample behavior of the procedures. The experiment-design details are given below.

4.1. TEST MODELS

(1) *Test set 1* (*AR models*). In this test we use five tailor-made AR processes to observe the performance of the procedure when the model assumptions are satisfied:

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

where $(\theta_1, \theta_2, \ldots, \theta_5) = (5.00, 5.25, 5.50, 5.75, 6{,}00)$, and the ε_{ij}'s are i.i.d. N(0, 1) random variables $\forall i, j$. The values of the θ's, which have no effect on the performance of the confidence-interval procedures, were chosen so that the test set could also be used in a companion study of a multiple-comparison procedure [21]. To avoid initial-condition bias, each model is initialized from its steady-state distribution.

(2) *Test set 2*. The models in this test set are identical to test set 1 except that the ε_{ij}'s are i.i.d. (shifted) exponentially distributed random variables with mean 0 and variance 1. The ε_{ij}'s are generated by first generating an observation from an exponential distribution with mean 1, and then subtracting 1 from the result. This set relaxes the normal residual assumption to observe the effect of the departure from normality. To initialize the models we set $Y_{ij} = \theta_i$ for $i = 1, 2, \ldots, 5$ and $j = -3, -2, -1, 0$, and then delete the initial 100 observations.

(3) *Test set 3* (*ARMA models*). In this test set we observe the effect of correlation structures different from AR processes by examining ARMA processes (ARMA processes are equivalent to infinite-order AR processes):

$$\begin{aligned}
Y_{1,j} - \theta_1 &= \varepsilon_{1,j} + 0.8\varepsilon_{1,j-1},\\
Y_{2,j} - \theta_2 &= 0.7(Y_{2,j-1} - \theta_2) + \varepsilon_{2,j} + 0.5\varepsilon_{2,j-1},\\
Y_{3,j} - \theta_3 &= 0.3(Y_{3,j-1} - \theta_3) + \varepsilon_{3,j} + 0.2\varepsilon_{3,j-1},\\
Y_{4,j} - \theta_4 &= 0.5(Y_{4,j-1} - \theta_4) + 0.25(Y_{4,j-2} - \theta_4) + \varepsilon_{4,j} + 0.3\varepsilon_{4,j-1},\\
Y_{5,j} - \theta_5 &= 0.4(Y_{5,j-1} - \theta_5) + 0.2(Y_{5,j-2} - \theta_5) + \varepsilon_{5,j} + 0.25\varepsilon_{5,j-1} + 0.1\varepsilon_{5,j-2},\\
Y_{6,j} - \theta_6 &= 0.35(Y_{6,j-1} - \theta_6) + 0.25(Y_{6,j-2} - \theta_6) + 0.15(Y_{6,j-3} - \theta_6) + \varepsilon_{6,j}\\
&\qquad + 0.2\varepsilon_{6,j-1},
\end{aligned}$$

where $(\theta_1, \theta_2, \ldots, \theta_6) = (4.4, 5.3, 5.0, 5.6, 5.9, 6.2)$, and the ε_{ij}'s are i.i.d. N(0,1) random variables $\forall i, j$. To initialize model 1, we set $Y_{1,1} = Z_1 + 0.8Z_2$, where Z_1 and Z_2 are independent N(0,1) random variables. To initialize the other models we set $Y_{i,j} = \theta_i$ and $\varepsilon_{ij} = 0$ for $i = 2, 3, \ldots, 6$ and $j = -2, -1, 0$, and then delete the initial 100 observations.

(4) *Test set 4* (*M/M/1 queue*). In this test set we study the performance of the confidence-interval procedures against a standard system-simulation example. Two parameter settings are selected: mean interarrival time 10 and mean service time 9 (implying a heavy traffic intensity of $\rho = 0.9$), and mean interarrival time 10 and mean service time 5 (implying a moderate traffic intensity of $\rho = 0.5$). For each queue we simultaneously record two processes: the system time and the number of entities in the system, where system time is the sum of wait and service time of a customer. The system-time process is indexed by consecutive arrivals. The number-of-entities process is the time-average number of entities in the system over consecutive observation intervals (every 40 time units). We sample the initial number of entities from the steady-state distribution $\beta_i = \rho^i(1 - \rho)$, for $i = 0, 1, 2, \ldots$, where β_i is the probability that i entities are initially in the system.

(5) *Test set 5* (*closed-queueing network*). In this test set we exercise the procedures on a system-simulation example with different characteristics. Consider

a production system that consists of 5 identical machines and s spares. Suppose 5 operators and 1 repair technician are in the system. When a machine breaks down it is immediately replaced by a spare machine if one is available. The broken machine is then sent to be repaired. The machine-failure times follow an exponential distribution with mean 10 time units. The repair time is also exponentially distributed with mean μ_2 time units. A repaired machine resumes production if there is an operator available; otherwise it is kept as a spare. For each interval of 40 time units we compute the time-average number of machines in operation. The output process is formed by the consecutive averages (i.e. batch means). We want to estimate the long-run expected number of operating machines for the 5 alternatives in table 1; the true, analytically-determined values are also shown in the table.

Table 1

(s, μ_2) combinations for the production system.

Alternative	s	μ_2	Expected number of operating machines	Expected number of time a machine is operating γ
1	4	3.25	3.033	0.367
2	3	3.00	3.218	0.438
3	2	2.50	3.604	0.561
4	1	2.00	3.891	0.686
5	0	1.75	3.772	0.754

We sample the initial number of machines in operation from a binomial distribution with parameters $(5 + s)$ and γ, where γ is given in the last column of table 1. To initialize each simulation we delete the initial 4000 time units.

4.2. EXPERIMENT FACTORS AND PERFORMANCE MEASURES

In the empirical study, two factors are controlled:

1. Degrees of freedom of $\widehat{\mathrm{Var}}[\bar{Y}]$. Three alternatives are used: Fishman's approximation, the equivalent sample size and the full sample size.

2. Sample size. We used three levels: $n = 500, 2500, 5000$ per system, except for the closed-queueing network (test set 5), where the sample sizes are $n = 200, 500, 1000$. The sample sizes were selected according to experience from pilot runs so that (we thought) the smallest value would be the minimum feasible for any procedure to work, and the largest value would be more than adequate.

To evaluate the performance of the confidence-interval procedures, we estimate coverage probability, expected halfwidth, and standard deviation of the confidence-interval halfwidth from 100 replications of the experiment (implying the first digit of the probability estimate is accurate, the second more uncertain). We separately compute these performance measures for cases that pass Portmanteau's test (with a significance level 0.1), and for all cases whether or not they pass the test. In all experiments the set of possible orders is $\mathcal{O} = \{1, 2, \ldots, 8\}$, where the maximum order was determined arbitrarily. The nominal confidence level is $1 - \alpha = 0.9$.

4.3. NUMERICAL RESULTS AND DISCUSSION

In this section we summarize the results of the experiments on each test set; numerical values are displayed in tables 2–25. General conclusions are given in section 5.

In test set 1, we find that the probability that Rissanen's order estimator selects the true order increases as the sample size increases, as expected (table 2). When it misses the true order the PLS estimator tends to overestimate it. The coverage of the confidence intervals is approximately the nominal level (table 3). The halfwidths look stable since the standard deviation of the halfwidth is quite small relative to the average halfwidth (tables 4,5). The three approximations for degrees of freedom do not make a noticeable difference in this case, and the performance of the intervals passing Portmanteau's test does not appear to be any better than the performance of all intervals when the test is ignored.

In test set 2 the overall results are similar to that of set 1, except that the estimated coverage is perhaps slightly lower when the sample size is small (refer to tables 6–9). No other degradation is apparent.

In test set 3, since there are moving-average terms and the true AR order is finite, the estimated orders tend to be larger (table 10). However, the confidence-interval performance is satisfactory: The nominal confidence level can be achieved in moderate or large sample sizes (table 11) and the halfwidths are stable (tables 12–13). Again, Portmanteau's test is not a good indicator of when the procedures will work.

In test set 4, when the traffic intensity is $\rho = 0.9$, the results for the system-time process show noticeable degradation: The coverage is low and the halfwidths tend to be wide (tables 14 –17). This is due to the strong positive correlation in the process. In particular, the halfwidths computed by using the equivalent sample size are unstable. When we increase the sample size the performance does not improve substantially, suggesting that even 5000 observations is not sufficient for this process. When the traffic intensity is 0.5, the correlation in the process decreases and the results improve. On the other hand, the results for the number of entities in the queue were insensitive to the change in traffic intensity (tables 18–21). In both cases the results were stable and satisfactory.

In test set 5, the results are similar to those for the number of entities in the *M*/*M*/1 experiments. The performance is satisfactory in terms of coverage and stability of the confidence intervals (tables 22–25). We conjecture that the process is well approximated by an AR model since the fitted orders are observed to be low (first or second order).

Table 2

Distribution of order estimates (set 1), where "total pass" means replications that passed the Portmanteau test.

model	sample	order								total
	size	1	2	3	4	5	6	7	8	pass
model 1	500	64	19	7	2	2	0	1	0	95
	2500	66	16	8	4	2	0	0	0	96
	5000	60	18	5	6	1	2	0	0	92
model 2	500	64	20	4	1	1	0	0	0	90
	2500	61	20	6	2	1	0	0	0	90
	5000	75	9	4	1	2	0	0	0	91
model 3	500	0	55	16	10	2	3	2	0	88
	2500	0	59	17	6	4	3	1	0	90
	5000	0	63	14	8	0	2	0	0	87
model 4	500	1	33	20	15	6	2	1	1	79
	2500	0	6	40	19	9	5	2	3	84
	5000	0	0	60	13	13	5	0	1	92
model 5	500	0	28	28	13	7	2	2	1	81
	2500	0	1	41	31	6	3	3	3	88
	5000	0	0	36	30	11	5	3	1	86

Table 3

Estimated coverage of 90% confidence intervals (set 1).

sample size		500		2500		5000	
model	df	pass only	overall	pass only	overall	pass only	overall
model 1	large sample	90.5	91.0	88.5	89.0	84.7	86.0
	Fishman	90.5	91.0	88.5	89.0	85.8	87.0
	equivalent sample	90.5	91.0	88.5	89.0	84.7	86.0
model 2	large sample	86.6	86.0	91.1	90.0	89.0	88.0
	Fishman	87.7	87.0	92.2	91.0	89.0	88.0
	equivalent sample	87.7	87.0	92.2	91.0	89.0	88.0
model 3	large sample	89.7	91.0	87.7	86.0	96.5	96.0
	Fishman	90.9	92.0	88.8	88.0	96.5	96.0
	equivalent sample	90.9	92.0	88.8	88.0	96.5	96.0
model 4	large sample	92.4	90.0	89.2	88.0	93.4	94.0
	Fishman	92.4	90.0	89.2	88.0	93.4	94.0
	equivalent sample	92.4	90.0	89.2	88.0	93.4	94.0
model 5	large sample	85.1	87.0	82.9	84.0	88.3	90.0
	Fishman	86.4	88.0	82.9	84.0	88.3	90.0
	equivalent sample	85.1	87.0	82.9	84.0	88.3	90.0

Table 4

Average halfwidth of 90% confidence intervals (set 1).

sample size		500		2500		5000	
model	df	pass only	overall	pass only	overall	pass only	overall
model 1	large sample	0.1038	0.1040	0.0469	0.0469	0.0332	0.0332
	Fishman	0.1047	0.1050	0.0470	0.0470	0.0332	0.0332
	equivalent sample	0.1040	0.1042	0.0469	0.0469	0.0332	0.0332
model 2	large sample	0.3660	0.3627	0.1649	0.1649	0.1160	0.1158
	Fishman	0.3795	0.3760	0.1661	0.1661	0.1164	0.1162
	equivalent sample	0.3840	0.3803	0.1664	0.1664	0.1166	0.1163
model 3	large sample	0.2916	0.2896	0.1304	0.1299	0.0921	0.0921
	Fishman	0.3041	0.3018	0.1314	0.1309	0.0924	0.0925
	equivalent sample	0.3007	0.2984	0.1311	0.1306	0.0923	0.0924
model 4	large sample	0.1759	0.1747	0.0814	0.0805	0.0578	0.0578
	Fishman	0.1811	0.1797	0.0819	0.0810	0.0580	0.0580
	equivalent sample	0.1775	0.1764	0.0815	0.0807	0.0579	0.0579
model 5	large sample	0.2676	0.2614	0.1287	0.1279	0.0903	0.0900
	Fishman	0.2804	0.2734	0.1300	0.1292	0.0908	0.0905
	equivalent sample	0.2746	0.2679	0.1294	0.1286	0.0906	0.0902

Table 5

Standard deviation of halfwidth of 90% confidence intervals (set 1).

sample size		500		2500		5000	
model	df	pass only	overall	pass only	overall	pass only	overall
model 1	large sample	0.0068	0.0070	0.0014	0.0014	0.0007	0.0007
	Fishman	0.0068	0.0070	0.0014	0.0014	0.0007	0.0007
	equivalent sample	0.0068	0.0070	0.0014	0.0014	0.0007	0.0007
model 2	large sample	0.0505	0.0504	0.0113	0.0113	0.0051	0.0052
	Fishman	0.0547	0.0546	0.0114	0.0114	0.0051	0.0052
	equivalent sample	0.0583	0.0580	0.0116	0.0116	0.0052	0.0052
model 3	large sample	0.0490	0.0479	0.0075	0.0075	0.0042	0.0042
	Fishman	0.0534	0.0522	0.0076	0.0076	0.0042	0.0042
	equivalent sample	0.0539	0.0526	0.0076	0.0076	0.0042	0.0042
model 4	large sample	0.0247	0.0264	0.0051	0.0054	0.0027	0.0027
	Fishman	0.0267	0.0283	0.0051	0.0055	0.0027	0.0027
	equivalent sample	0.0255	0.0272	0.0051	0.0054	0.0027	0.0027
model 5	large sample	0.0503	0.0488	0.0108	0.0116	0.0054	0.0055
	Fishman	0.0568	0.0550	0.0111	0.0119	0.0055	0.0056
	equivalent sample	0.0553	0.0535	0.0110	0.0118	0.0055	0.0055

Table 6

Distribution of order estimates (set 2), where "total pass" means replications that passed the Portmanteau test.

model	sample	order								total
	size	1	2	3	4	5	6	7	8	pass
model 1	500	56	15	8	3	2	2	2	2	90
	2500	70	11	4	3	3	1	0	0	92
	5000	65	12	5	2	2	1	1	0	88
model 2	500	68	11	9	3	2	0	1	0	94
	2500	62	8	4	5	1	0	1	2	83
	5000	70	8	8	1	2	0	0	0	89
model 3	500	1	50	20	8	6	3	1	1	90
	2500	0	59	22	4	4	3	2	4	98
	5000	0	65	11	2	5	4	2	2	91
model 4	500	0	35	21	7	9	1	1	3	77
	2500	0	5	50	11	6	1	7	4	84
	5000	0	0	55	20	6	4	6	3	94
model 5	500	0	30	17	15	5	3	4	5	79
	2500	0	2	36	25	7	5	3	2	80
	5000	0	0	33	31	9	9	2	3	87

Table 7

Estimated coverage of 90% confidence intervals (set 2).

sample size		500		2500		5000	
model	df	pass only	overall	pass only	overall	pass only	overall
model 1	large sample	86.6	88.0	94.5	94.0	95.4	95.0
	Fishman	87.7	89.0	94.5	94.0	95.4	95.0
	equivalent sample	86.6	88.0	94.5	94.0	95.4	95.0
model 2	large sample	73.4	75.0	85.5	86.0	94.3	94.0
	Fishman	74.4	76.0	85.5	86.0	94.3	94.0
	equivalent sample	75.5	77.0	85.5	86.0	94.3	94.0
model 3	large sample	92.2	91.0	86.7	87.0	89.0	90.0
	Fishman	92.2	91.0	86.7	87.0	89.0	90.0
	equivalent sample	92.2	91.0	86.7	87.0	89.0	90.0
model 4	large sample	87.0	90.0	89.2	88.0	87.2	88.0
	Fishman	87.0	90.0	89.2	88.0	87.2	88.0
	equivalent sample	87.0	90.0	89.2	88.0	87.2	88.0
model 5	large sample	84.8	84.0	93.7	93.0	88.5	89.0
	Fishman	84.8	84.0	93.7	93.0	88.5	89.0
	equivalent sample	84.8	84.0	93.7	93.0	88.5	89.0

Table 8

Average halfwidth of 90% confidence intervals (set 2).

sample size		500		2500		5000	
model	df	pass only	overall	pass only	overall	pass only	overall
model 1	large sample	0.1056	0.1058	0.0467	0.0466	0.0332	0.0332
	Fishman	0.1067	0.1069	0.0467	0.0467	0.0333	0.0332
	equivalent sample	0.1058	0.1060	0.0467	0.0466	0.0332	0.0332
model 2	large sample	0.3540	0.3541	0.1613	0.1623	0.1163	0.1165
	Fishman	0.3667	0.3668	0.1625	0.1635	0.1167	0.1169
	equivalent sample	0.3702	0.3702	0.1627	0.1638	0.1168	0.1170
model 3	large sample	0.2855	0.2802	0.1310	0.1308	0.0929	0.0931
	Fishman	0.2973	0.2913	0.1321	0.1319	0.0933	0.0935
	equivalent sample	0.2936	0.2879	0.1318	0.1316	0.0932	0.0934
model 4	large sample	0.1744	0.1695	0.0814	0.0805	0.0580	0.0578
	Fishman	0.1795	0.1743	0.0819	0.0820	0.0582	0.0580
	equivalent sample	0.1760	0.1711	0.0815	0.0806	0.0580	0.0578
model 5	large sample	0.2646	0.2599	0.1273	0.1256	0.0913	0.0909
	Fishman	0.2774	0.2719	0.1286	0.1269	0.0918	0.0914
	equivalent sample	0.2712	0.2662	0.1280	0.1263	0.0915	0.0912

Table 9

Standard deviation of halfwidth of 90% confidence intervals (set 2).

sample size		500		2500		5000	
model	df	pass only	overall	pass only	overall	pass only	overall
model 1	large sample	0.0090	0.0093	0.0020	0.0020	0.0010	0.0010
	Fishman	0.0093	0.0096	0.0020	0.0020	0.0010	0.0010
	equivalent sample	0.0091	0.0094	0.0020	0.0020	0.0010	0.0010
model 2	large sample	0.0551	0.0554	0.0117	0.0118	0.0053	0.0052
	Fishman	0.0588	0.0592	0.0118	0.0119	0.0053	0.0052
	equivalent sample	0.0620	0.0624	0.0120	0.0121	0.0053	0.0052
model 3	large sample	0.0457	0.0474	0.0094	0.0093	0.0044	0.0044
	Fishman	0.0492	0.0512	0.0095	0.0095	0.0044	0.0044
	equivalent sample	0.0491	0.0509	0.0095	0.0095	0.0044	0.0044
model 4	large sample	0.0241	0.0257	0.0050	0.0061	0.0030	0.0031
	Fishman	0.0259	0.0276	0.0051	0.0062	0.0030	0.0031
	equivalent sample	0.0248	0.0263	0.0051	0.0061	0.0030	0.0031
model 5	large sample	0.0441	0.0478	0.0116	0.0119	0.0064	0.0069
	Fishman	0.0494	0.0533	0.0119	0.0122	0.0065	0.0070
	equivalent sample	0.0474	0.0512	0.0118	0.0121	0.0065	0.0070

Table 10

Distribution of order estimates (set 3), where "total pass" means replications that passed the Portmanteau test.

model	sample size	order 1	2	3	4	5	6	7	8	total pass
model 1	500	0	0	0	11	12	17	11	8	59
	2500	0	0	0	0	0	3	2	19	24
	5000	0	0	0	0	0	0	1	15	16
model 2	500	0	11	26	19	7	6	3	3	75
	2500	0	0	11	36	17	14	3	4	85
	5000	0	0	1	27	28	16	4	4	80
model 3	500	41	25	9	5	1	0	0	0	81
	2500	13	54	14	6	4	2	0	0	93
	5000	1	56	20	8	4	1	1	0	91
model 4	500	60	17	7	1	3	2	0	0	90
	2500	73	12	5	0	0	0	1	0	91
	5000	62	20	5	3	1	1	0	1	93
model 5	500	40	18	9	7	2	2	1	0	79
	2500	6	11	37	15	7	2	1	1	80
	5000	0	1	52	17	7	6	1	1	85
model 6	500	3	31	26	10	5	5	0	1	81
	2500	0	1	55	25	6	3	2	0	92
	5000	0	0	55	19	10	5	0	2	91

Table 11

Estimated coverage of 90% confidence intervals (set 3).

sample size		500		2500		5000	
model	df	pass only	overall	pass only	overall	pass only	overall
model 1	large sample	88.1	88.0	79.1	85.0	81.2	84.0
	Fishman	88.1	88.0	79.1	87.0	81.2	84.0
	equivalent sample	88.1	88.0	79.1	85.0	81.2	84.0
model 2	large sample	83.9	85.0	91.7	93.0	89.9	90.0
	Fishman	83.9	85.0	91.7	93.0	89.9	90.0
	equivalent sample	83.9	85.0	92.9	94.0	89.9	90.0
model 3	large sample	87.6	89.0	92.4	93.0	91.2	92.0
	Fishman	87.6	89.0	93.5	94.0	91.2	92.0
	equivalent sample	87.6	89.0	93.5	94.0	91.2	92.0
model 4	large sample	88.8	87.0	89.0	89.0	89.2	89.0
	Fishman	89.9	88.0	89.0	89.0	89.2	89.0
	equivalent sample	89.9	90.0	90.1	90.0	89.2	89.0
model 5	large sample	92.4	91.0	91.2	90.0	87.0	88.0
	Fishman	94.9	93.0	91.2	90.0	87.0	88.0
	equivalent sample	94.9	93.0	91.2	90.0	87.0	88.0
model 6	large sample	80.2	82.0	92.3	93.0	90.1	90.0
	Fishman	81.4	83.0	93.4	94.0	90.1	90.0
	equivalent sample	80.2	82.0	92.3	93.0	90.1	90.0

Table 12

Average halfwidth of 90% confidence intervals (set 3).

sample size		500		2500		5000	
model	df	pass only	overall	pass only	overall	pass only	overall
model 1	large sample	0.1266	0.1289	0.0567	0.0580	0.0407	0.0410
	Fishman	0.1298	0.1319	0.0571	0.0584	0.0408	0.0412
	equivalent sample	0.1271	0.1295	0.0568	0.0581	0.0407	0.0411
model 2	large sample	0.3612	0.3600	0.1632	0.1638	0.1163	0.1164
	Fishman	0.3743	0.3725	0.1644	0.1650	0.1167	0.1168
	equivalent sample	0.3789	0.3775	0.1647	0.1653	0.1168	0.1169
model 3	large sample	0.1285	0.1293	0.0562	0.0565	0.0396	0.0397
	Fishman	0.1301	0.1308	0.0563	0.0566	0.0396	0.0397
	equivalent sample	0.1291	0.1298	0.0562	0.0565	0.0396	0.0397
model 4	large sample	0.3614	0.3622	0.1700	0.1706	0.1209	0.1209
	Fishman	0.3746	0.3755	0.1713	0.1719	0.1214	0.1213
	equivalent sample	0.3784	0.3795	0.1717	0.1732	0.1215	0.1215
model 5	large sample	0.2505	0.2494	0.1106	0.1113	0.0775	0.0779
	Fishman	0.2572	0.2559	0.1112	0.1119	0.0777	0.0781
	equivalent sample	0.2557	0.2546	0.1110	0.1117	0.0776	0.0781
model 6	large sample	0.3323	0.3257	0.1572	0.1567	0.1130	0.1128
	Fishman	0.3493	0.3418	0.1589	0.1584	0.1136	0.1134
	equivalent sample	0.3460	0.3386	0.1585	0.1580	0.1135	0.1132

Table 13

Standard deviation of halfwidth of 90% confidence intervals (set 3).

sample size		500		2500		5000	
model	df	pass only	overall	pass only	overall	pass only	overall
model 1	large sample	0.0157	0.0180	0.0034	0.0043	0.0014	0.0021
	Fishman	0.0166	0.0189	0.0035	0.0043	0.0014	0.0021
	equivalent sample	0.0159	0.0182	0.0034	0.0043	0.0014	0.0021
model 2	large sample	0.0561	0.0563	0.0113	0.0110	0.0063	0.0059
	Fishman	0.0610	0.0611	0.0114	0.0111	0.0064	0.0059
	equivalent sample	0.0649	0.0649	0.0116	0.0113	0.0064	0.0060
model 3	large sample	0.0126	0.0124	0.0030	0.0032	0.0012	0.0013
	Fishman	0.0128	0.0126	0.0030	0.0032	0.0013	0.0013
	equivalent sample	0.0128	0.0126	0.0030	0.0032	0.0012	0.0013
model 4	large sample	0.0468	0.0503	0.0107	0.0111	0.0047	0.0046
	Fishman	0.0508	0.0547	0.0109	0.0113	0.0047	0.0046
	equivalent sample	0.0541	0.0585	0.0110	0.0114	0.0047	0.0046
model 5	large sample	0.0323	0.0327	0.0071	0.0072	0.0034	0.0038
	Fishman	0.0343	0.0348	0.0072	0.0072	0.0035	0.0038
	equivalent sample	0.0346	0.0350	0.0072	0.0072	0.0034	0.0038
model 6	large sample	0.0590	0.0599	0.0121	0.0127	0.0058	0.0061
	Fishman	0.0661	0.0671	0.0123	0.0130	0.0059	0.0062
	equivalent sample	0.0667	0.0674	0.0124	0.0130	0.0059	0.0062

Table 14

Distribution of order estimates (set 4, system time), where "total pass" means replications that passed the Portmanteau test.

traffic	sample	order								total
intensity	size	1	2	3	4	5	6	7	8	pass
$\rho = 0.9$	500	57	17	5	3	3	2	2	1	90
	2500	58	11	5	3	2	1	2	1	83
	5000	59	16	13	2	4	2	0	0	96
$\rho = 0.5$	500	58	19	4	4	3	0	2	2	92
	2500	49	16	5	1	1	2	0	2	76
	5000	37	30	9	3	1	0	0	0	80

Table 15

Estimated coverage of 90% confidence intervals (set 4, system time).

traffic intensity	sample size	500		2500		5000	
	df	pass only	overall	pass only	overall	pass only	overall
$\rho = 0.9$	large sample	56.6	56.0	73.4	76.0	64.5	66.0
	Fishman	59.9	59.0	75.9	78.0	67.7	69.0
	equivalent sample	83.3	80.0	90.3	91.0	84.3	85.0
$\rho = 0.5$	large sample	84.7	84.0	84.2	83.0	86.2	87.0
	Fishman	84.7	84.0	84.2	83.0	86.2	87.0
	equivalent sample	84.7	84.0	84.2	83.0	86.2	87.0

Table 16

Average halfwidth of 90% confidence intervals (set 4, system time).

traffic intensity	sample size	500		2500		5000	
	df	pass only	overall	pass only	overall	pass only	overall
$\rho = 0.9$	large sample	62.58	99.98	36.82	38.32	24.36	25.00
	Fishman	165.74	311.32	53.30	55.37	27.31	28.15
	equivalent sample	233.22	376.11	134.29	140.32	80.66	83.37
$\rho = 0.5$	large sample	1.806	1.857	0.868	0.888	0.631	0.634
	Fishman	1.862	1.917	0.873	0.893	0.633	0.636
	equivalent sample	1.868	1.932	0.874	0.894	0.633	0.636

Table 17

Standard deviation of halfwidth of 90% confidence intervals (set 4, system time).

traffic intensity	sample size	500		2500		5000	
	df	pass only	overall	pass only	overall	pass only	overall
$\rho = 0.9$	large sample	231.01	453.15	33.62	34.62	18.86	19.71
	Fishman	892.34	1745.79	114.41	109.38	29.60	30.44
	equivalent sample	886.58	1737.59	133.88	137.66	80.24	83.40
$\rho = 0.5$	large sample	0.613	0.704	0.183	0.193	0.099	0.096
	Fishman	0.654	0.755	0.186	0.196	0.100	0.097
	equivalent sample	0.684	0.814	0.187	0.197	0.100	0.097

Table 18

Distribution of order estimates (set 4, number of entities), where "total pass" means replications that passed the Portmanteau test.

traffic	sample	order								total
intensity	size	1	2	3	4	5	6	7	8	pass
$\rho = 0.9$	500	4	46	20	8	4	0	0	0	82
	2500	0	22	37	11	3	1	3	1	78
	5000	0	4	53	14	5	4	1	1	82
$\rho = 0.5$	500	46	18	8	11	1	3	1	2	90
	2500	29	34	6	3	3	1	5	2	83
	5000	5	48	13	2	6	5	1	3	83

Table 19

Estimated coverage of 90% confidence intervals (set 4, number of entities).

traffic	sample size	500		2500		5000	
intensity	df	pass only	overall	pass only	overall	pass only	overall
$\rho = 0.9$	large sample	67.0	68.0	80.7	80.0	86.5	86.0
	Fishman	71.9	74.0	80.7	80.0	87.8	87.0
	equivalent sample	75.6	78.0	82.0	81.0	92.6	92.0
$\rho = 0.5$	large sample	91.1	91.0	91.5	92.0	90.3	89.0
	Fishman	91.1	92.0	91.5	92.0	90.3	89.0
	equivalent sample	91.1	91.0	91.5	92.0	90.3	89.0

Table 20

Average halfwidth of 90% confidence intervals (set 4, number of entities).

traffic	sample size	500		2500		5000	
intensity	df	pass only	overall	pass only	overall	pass only	overall
$\rho = 0.9$	large sample	4.63	4.44	2.01	2.09	1.73	1.73
	Fishman	9.47	8.55	2.10	2.16	1.78	1.78
	equivalent sample	15.24	14.55	3.88	4.11	2.82	2.85
$\rho = 0.5$	large sample	0.1232	0.1242	0.0571	0.0574	0.0400	0.0405
	Fishman	0.1249	0.1259	0.0573	0.0575	0.0401	0.0406
	equivalent sample	0.1238	0.1247	0.0572	0.0574	0.0400	0.0405

Table 21

Standard deviation of halfwidth of 90% confidence intervals (set 4, number of entities).

traffic	sample size	500		2500		5000	
intensity	df	pass only	overall	pass only	overall	pass only	overall
$\rho = 0.9$	large sample	7.73	7.03	0.92	0.95	0.66	0.66
	Fishman	29.58	26.84	1.03	1.06	0.70	0.71
	equivalent sample	30.52	27.84	4.17	4.33	2.68	2.77
$\rho = 0.5$	large sample	0.0323	0.0320	0.0071	0.0069	0.0037	0.0039
	Fishman	0.0331	0.0328	0.0072	0.0070	0.0037	0.0039
	equivalent sample	0.0329	0.0325	0.0071	0.0070	0.0037	0.0039

Table 22

Distribution of order estimates (set 5), where "total pass" means replications that passed the Portmanteau test.

model	sample size	order 1	2	3	4	5	6	7	8	total pass
alternative 1	200	51	19	11	7	2	1	1	0	92
	500	51	21	10	4	0	1	0	1	88
	1000	51	27	9	3	0	0	0	1	91
alternative 2	200	58	24	8	0	0	3	0	0	93
	500	60	20	4	5	1	0	0	0	90
	1000	54	26	11	1	1	0	0	0	93
alternative 3	200	62	17	7	0	2	0	0	1	89
	500	58	17	5	6	2	0	1	0	89
	1000	60	23	4	2	1	1	0	0	91
alternative 4	200	63	19	6	4	2	1	0	0	95
	500	64	14	7	2	0	1	0	1	86
	1000	59	14	10	4	2	0	1	0	90
alternative 5	200	63	11	8	1	3	0	1	0	87
	500	46	24	4	4	2	2	2	1	85
	1000	59	19	9	2	1	3	0	0	93

Table 23

Estimated coverage of 90% confidence intervals (set 5).

sample size		200		500		1000	
model	df	pass only	overall	pass only	overall	pass only	overall
alternative 1	large sample	85.8	87.0	92.0	92.0	92.3	92.0
	Fishman	86.9	88.0	92.0	92.0	92.3	92.0
	equivalent sample	85.8	87.0	92.0	92.0	92.3	92.0
alternative 2	large sample	91.3	91.0	91.1	90.0	87.0	87.0
	Fishman	91.3	91.0	91.1	90.0	87.0	87.0
	equivalent sample	91.3	91.0	91.1	90.0	87.0	87.0
alternative 3	large sample	87.6	89.0	83.1	85.0	85.7	87.0
	Fishman	88.7	90.0	83.1	85.0	85.7	87.0
	equivalent sample	87.6	89.0	83.1	85.0	85.7	87.0
alternative 4	large sample	87.3	88.0	84.8	84.0	89.9	88.0
	Fishman	87.3	88.0	86.0	86.0	89.9	88.0
	equivalent sample	87.3	88.0	84.8	84.0	89.9	88.0
alternative 5	large sample	89.6	90.0	89.4	91.0	91.3	91.0
	Fishman	89.6	90.0	90.5	92.0	91.3	91.0
	equivalent sample	89.6	90.0	89.4	91.0	91.3	91.0

Table 24

Average halfwidth of 90% confidence intervals (set 5).

sample size		200		500		1000	
model	df	pass only	overall	pass only	overall	pass only	overall
alternative 1	large sample	0.1326	0.1332	0.0855	0.0857	0.0598	0.0599
	Fishman	0.1356	0.1361	0.0862	0.0864	0.0600	0.0601
	equivalent sample	0.1329	0.1335	0.0855	0.0857	0.0598	0.0599
alternative 2	large sample	0.1271	0.1273	0.0814	0.0811	0.0570	0.0570
	Fishman	0.1296	0.1298	0.0820	0.0817	0.0572	0.0572
	equivalent sample	0.1274	0.1276	0.0815	0.0812	0.0570	0.0571
alternative 3	large sample	0.1095	0.1094	0.0678	0.0677	0.0482	0.0483
	Fishman	0.1115	0.1112	0.0683	0.0682	0.0483	0.0485
	equivalent sample	0.1109	0.1096	0.0678	0.0677	0.0482	0.0483
alternative 4	large sample	0.0797	0.0798	0.0503	0.0504	0.0359	0.0359
	Fishman	0.0811	0.0813	0.0506	0.0507	0.0360	0.0360
	equivalent sample	0.0798	0.0799	0.0503	0.0504	0.0359	0.0359
alternative 5	large sample	0.0611	0.0607	0.0386	0.0387	0.0277	0.0276
	Fishman	0.0620	0.0616	0.0388	0.0390	0.0277	0.0277
	equivalent sample	0.0611	0.0607	0.0386	0.0387	0.0277	0.0276

Table 25

Standard deviation of halfwidth of 90% confidence intervals (set 5).

sample size		200		500		1000	
model	df	pass only	overall	pass only	overall	pass only	overall
alternative 1	large sample	0.0179	0.0177	0.0063	0.0062	0.0028	0.0028
	Fishman	0.0193	0.0190	0.0065	0.0064	0.0028	0.0028
	equivalent sample	0.0182	0.0180	0.0063	0.0062	0.0028	0.0028
alternative 2	large sample	0.0155	0.0155	0.0060	0.0059	0.0030	0.0030
	Fishman	0.0168	0.0167	0.0061	0.0060	0.0030	0.0030
	equivalent sample	0.0157	0.0157	0.0060	0.0059	0.0030	0.0030
alternative 3	large sample	0.0114	0.0113	0.0049	0.0049	0.0023	0.0024
	Fishman	0.0119	0.0118	0.0050	0.0049	0.0023	0.0024
	equivalent sample	0.0115	0.0114	0.0049	0.0049	0.0023	0.0024
alternative 4	large sample	0.0103	0.0102	0.0032	0.0031	0.0019	0.0018
	Fishman	0.0109	0.0108	0.0032	0.0031	0.0019	0.0018
	equivalent sample	0.0104	0.0103	0.0032	0.0032	0.0019	0.0018
alternative 5	large sample	0.0068	0.0069	0.0030	0.0034	0.0015	0.0015
	Fishman	0.0069	0.0070	0.0030	0.0035	0.0015	0.0015
	equivalent sample	0.0068	0.0069	0.0030	0.0034	0.0015	0.0015

5. Conclusions and recommendations

Based on the empirical results in section 4.3, a number of conclusions can be drawn:

- The AR-confidence-interval procedures perform well if an output process is not too strongly positively correlated. With this or any method, one needs to be cautious in the presence of strong positive correlations, in which case a large sample is required to prevent the procedure from degrading. It is a good practice to compute sample autocorrelations, which help users to predict the performance of the procedure based on the correlation strengths.
- PLS tends to overestimate the order, but this is actually comforting. Overestimation is less harmful than underestimation because an AR(p) process can be treated as an AR($p + h$) process with h zero coefficients. Given the typically good performance of the confidence-interval procedures we conjecture that PLS is appropriately adjusting for the correlation structure in the output process at hand.
- The degrees of freedom for $\widehat{\mathrm{Var}}[\bar{Y}]$ affect the t critical values used in computing confidence intervals. Although we did not observe any substantial difference among the three procedures in most experiments, the equivalent sample size procedure did provide better coverage in the $M/M/1$ experiments on system time (table 15), which was the most difficult case for our AR procedure. Therefore we recommend using the equivalent sample size to determine the degrees of freedom.
- Portmanteau's test does not provide protection for the procedures. We conjecture that $\widehat{\mathrm{Var}}[\bar{Y}]$ based on an AR representation is a good approximation for $\widehat{\mathrm{Var}}[\bar{Y}]$ even when the actual output process is not AR(p).
- The procedures seem robust to deviation from normality or non-AR correlation structure when the sample size is moderate or large. The explanation could be the same as in the previous paragraph: $\widehat{\mathrm{Var}}[\bar{Y}]$ serves as a good approximation to $\mathrm{Var}[\bar{Y}]$ regardless of whether or not the output process is truly AR.

The empirical results also suggest future research. Since the procedure works well on moderately-correlated processes, and the correlations in a time series can be reduced by batching the process, improvements are expected by implementing the procedure on the batch means as we did for the number-of-entities process in test sets 4 and 5. This could lead to a generalization of the nonoverlapping-batch-means method. Some advantages over conventional batching strategies are possible. A central issue is how to determine a batch size that reduces the correlations while retaining enough data to estimate the AR model.

Appendix

ALGORITHM FOR COMPUTING PLS_h

Our algorithm is based on the following well-known lemma (see [15, p. 459]):

LEMMA 5.1

Let $\mathbf{A}$ be an $n \times n$ nonsingular matrix and $\boldsymbol{d}$ be an $n \times 1$ vector. Then

$$(\mathbf{A} + \boldsymbol{dd}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\boldsymbol{dd}'\mathbf{A}^{-1}}{1 + \boldsymbol{d}'\mathbf{A}^{-1}\boldsymbol{d}}.$$

Since

$$\mathbf{X}(i+1, h) = \begin{bmatrix} \mathbf{X}(i,h) \\ (1, Y_i, Y_{i-1}, \ldots, Y_{i-h+1}) \end{bmatrix} = \begin{bmatrix} \mathbf{X}(i,h) \\ \boldsymbol{d}' \end{bmatrix},$$

we can write $\mathbf{X}'(i+1,h)\mathbf{X}(i+1,h) = \mathbf{X}'(i,h)\mathbf{X}(i,h) + \boldsymbol{dd}'$. Therefore, by lemma 5.1,

$$[\mathbf{X}'(i+1,h)\mathbf{X}(i+1,h)]^{-1} = [\mathbf{X}'(i,h)\mathbf{X}(i,h)]^{-1} + \frac{[\mathbf{X}'(i,h)\mathbf{X}(i,h)]^{-1}\boldsymbol{dd}'[\mathbf{X}'(i,h)\mathbf{X}(i,h)]^{-1}}{1 + \boldsymbol{d}'[\mathbf{X}'(i,h)\mathbf{X}(i,h)]^{-1}\boldsymbol{d}}.$$

Thus, the updated matrix can be inverted without direct evaluation. From our experience, the rounding error of recursively updating is negligible when double precision is used. An algorithm for computing PLS_h is as follows:

```
Step 1. $l \leftarrow 2h + 1$.
Step 2. Compute $\mathbf{X}'(l,h)\mathbf{X}(l,h)$.
Step 3. Invert $\mathbf{X}'(l,h)\mathbf{X}(l,h)$.
Step 4. Compute $e_{l+1}(h)$ and set sum $\leftarrow e_{l+1}(h)$.
Step 5. for $i \leftarrow l+1$ to $n-1$, do
            update $[\mathbf{X}'(i,h)\mathbf{X}(i,h)]^{-1}$ from $[\mathbf{X}'(i-1,h)\mathbf{X}(i-1,h)]^{-1}$ using lemma 5.1
            compute $e_{i+1}(h)$
            sum $\leftarrow$ sum + $e_{i+1}(h)$
        endfor
Step 6. $\mathrm{PLS}_h = \mathrm{sum}/(n-l)$
```

There is a possibility that the matrix computed in step 2 is singular. In that case we may initialize l in step 1 with a larger integer such that $\mathbf{X}'(l,h)\mathbf{X}(l,h)$ is invertible. Also recall that $\hat{\phi}(n,h) = [\mathbf{X}'(n,h)\mathbf{X}(n,h)]^{-1}\mathbf{X}'(n,h)\mathbf{Y}(n,h)$. The matrix $[\mathbf{X}'(n,h)\mathbf{X}(n,h)]^{-1}$ can be obtained by updating $\mathbf{X}'(n-1,h)\mathbf{X}(n-1,h)$, which is available after we evaluate $e_n(h)$. This shows that the AR coefficient estimator is a by-product of the PLS order identification.

Proof of lemma 3.3

For any $\varepsilon > 0$,

$$\Pr\{|\hat{\phi}_i(n,p) - \phi_i| < \varepsilon\} = \Pr\{|\hat{\phi}_i(n,p) - \phi_i| < \varepsilon, \hat{p} = p\} + \sum_{j \in \mathcal{O}, j \neq p} \Pr\{|\hat{\phi}_i(n,p) - \phi_i| < \varepsilon, \hat{p} = j\},$$

where $\mathcal{O}$ is a set of possible orders. Since $\Pr\{|\hat{\phi}_i(n,p) - \phi_i| < \varepsilon, \hat{p} = j\} \leq \Pr\{\hat{p} = j\}$, it follows that

$$\lim_{n\to\infty} \Pr\{|\hat{\phi}_i(n,p) - \phi_i| \leq \varepsilon, \hat{p} = j\} \leq \lim_{n\to\infty} \Pr\{\hat{p} = j\} = 0, \quad \forall j \neq p.$$

Thus, $\lim_{n\to\infty} \Pr\{|\hat{\phi}_i(n,p) - \phi_i| < \varepsilon\} = \lim_{n\to\infty} \Pr\{|\hat{\phi}_i(n,p) - \phi_i| < \varepsilon \text{ and } \hat{p} = p\} + 0 = 1$, $\forall i$, from lemma 3.2. □

Proof of theorem 3.1

Since $\hat{p} \to p$ with probability 1, $\lim_{n\to\infty} \hat{\phi}(n,\hat{p})$ is a p vector. To prove the theorem, we need to show that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\{|\hat{\phi}_i(n,\hat{p}) - \phi_i| < \varepsilon\} = 1, \quad \text{for } i = 1, 2, \ldots, p.$$

Notice that

$$\Pr\{|\hat{\phi}_i(n,\hat{p}) - \phi_i| < \varepsilon\} = \Pr\{|\hat{\phi}_i(n,\hat{p}) - \phi_i| < \varepsilon, \hat{p} = p\} + \sum_{j \in \mathcal{O}, j \neq p} \Pr\{|\hat{\phi}_i(n,\hat{p}) - \phi_i| < \varepsilon, \hat{p} = j\},$$

for $i = 1, 2, \ldots, p$. Then from lemma 3.3,

$$\lim_{n\to\infty} \Pr\{|\hat{\phi}_i(n,\hat{p}) - \phi_i| < \varepsilon, \hat{p} = p\} = \lim_{n\to\infty} \Pr\{|\hat{\phi}_i(n,p) - \phi_i| < \varepsilon, \hat{p} = p\} = 1, \quad \forall i.$$

It follows that $\lim_{n\to\infty} \Pr\{|\hat{\phi}_i(n,\hat{p}) - \phi_i| < \varepsilon\} = 1$, $\forall i$. □

Proof of theorem 3.2

Let

$$S_n^2 = \frac{\sum_{j=p+1}^{n}\left(Y_j - \phi_0 - \sum_{k=1}^{p} \phi_k Y_{j-k}\right)^2}{n-p},$$

$$\hat{S}_n^2 = \frac{\sum_{j=p+1}^{n}\left(Y_j - \hat{\phi}_0(n,p) - \sum_{k=1}^{p} \hat{\phi}_k(n,p) Y_{j-k}\right)^2}{n-p}.$$

Notice that $\lim_{n\to\infty} S_n^2 = \sigma^2$ with probability 1 from the strong law of large numbers. The proof is completed by showing that $\hat{S}_n^2 - S_n^2 \xrightarrow{\mathcal{P}} 0$ and $\hat{\sigma}^2 - \hat{S}_n^2 \xrightarrow{\mathcal{P}} 0$. By adding and subtracting ϕ_i appropriately we can write $\hat{S}_n^2 = S_n^2 + R_n$, where $R_n \xrightarrow{\mathcal{P}} 0$, which establishes the first part. For any $\varepsilon > 0$,

$$\begin{aligned}\lim_{n\to\infty} \Pr\{|\hat{\sigma}^2 - \hat{S}_n^2| > \varepsilon\} &= \lim_{n\to\infty}\Big(\Pr\{|\hat{\sigma}^2 - \hat{S}_n^2| > \varepsilon, \hat{p} = p\} \\ &\quad + \Pr\{|\hat{\sigma}^2 - \hat{S}_n^2| > \varepsilon, \hat{p} \neq p\}\Big) \\ &\leq \lim_{n\to\infty}\Big(\Pr\{|\hat{S}_n^2 - \hat{S}_n^2| > \varepsilon, \hat{p} = p\} + \Pr\{\hat{p} \neq p\}\Big) \\ &= 0,\end{aligned}$$

where the final equality follows from the fact that $\hat{\sigma}^2 = \hat{S}_n^2$ when $\hat{p} = p$, and $\hat{p}$ is consistent for p. This completes the proof. □

Proof of theorem 3.3

Since $\hat{\sigma}^2 \xrightarrow{\mathcal{P}} \sigma^2$ (by theorem 3.2), and $\sum_{i=1}^{\hat{p}} \hat{\phi}(n,\hat{p}) \xrightarrow{\mathcal{P}} \sum_{i=1}^{p} \phi_i$ (by corollary 3.1), we have $n\widehat{\mathrm{Var}}[\bar{Y}] \xrightarrow{\mathcal{P}} \sigma^2/(1 - \sum_{i=1}^{p} \phi_i)^2 = \lim_{n\to\infty} n\,\mathrm{Var}[\bar{Y}]$. □

Proof of theorem 3.4

The conclusion that $(\bar{Y} - \theta)/\sqrt{\tau^2/n} \Rightarrow \mathrm{N}(0,1)$ follows from appendix 2 of [13] and theorem 21.1 of [2]. Then since $n\widehat{\mathrm{Var}}[\bar{Y}] \xrightarrow{\mathcal{P}} \tau^2$, by theorem 3.3, the result follows from Slutsky's Theorem. □

Acknowledgements

This material is based upon work supported by National Science Foundation Grant No. DDM-8922721. The authors acknowledge the valuable comments and suggestions from James R. Wilson and three anonymous referees.

References

[1] H. Akaike, Maximum likelihood identification of Gaussian auto-regressive moving average models, Biometrika 60(1973)255–266.

[2] P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).

[3] G.E.P. Box and G.M. Jenkins, *Time Series Analysis*, rev. ed. (Holden–Day, San Francisco, CA, 1976).

[4] Y. Chun, Time-series models of batch-means processes in simulation analysis, Working Paper 89-5, Department of Operations and Management Science, University of Minnesota (1989).

[5] J.D. Cryer, *Time Series Analysis* (Duxbury Press, Boston, 1986).

[6] N. Davies and P. Newbold, Some power studies of a Portmanteau test of time series models specification, Biometrika 66(1979)153–155.

[7] G.S. Fishman, *Concepts and Methods in Discrete Event Digital Simulation* (Wiley, New York, 1973).

[8] G.S. Fishman, *Principles of Discrete Event Simulation* (Wiley, New York, 1978).

[9] H.L. Gray, G.D. Kelly and D.D. McIntire, A new approach to ARMA modelling, Commun. Statist. B7(1978)1–77.

[10] E.J. Hannan, *Multiple Time Series* (Wiley, New York, 1970).

[11] E.J. Hannan, A.J. McDougall and D.S. Poskitt, Recursive estimation of autoregressions, J. Roy. Statist. Soc. Ser. B51(1989)217–233.

[12] E.M. Hemerly and M.H.A. Davis, Strong consistency of the PLS criterion for order determination of autoregressive processes, Ann. Statist. 17(1989)941–946.

[13] R.A. Johnson and M. Bagshaw, The effect of serial correlation on the performance of CUSUM tests, Technometrics 16(1974)103–112.

[14] G.M. Ljung and G.E.P. Box, On a measure of lack of fit in time series models, Biometrika 65(1978)67–72.

[15] K.V. Mardia, J.T. Kent and J.M. Bibby, *Multivariate Analysis* (Academic Press, New York, 1979).

[16] J. Rissanen, Order estimation by accumulated prediction errors, in: *Essays in Time Series and Allied Processes*, ed. J. Gani and M.B. Priestly, J. Appl. Prob. 23A(1986)55–61.

[17] T.J. Schriber and R.W. Andrews, ARMA-based confidence intervals for simulation output analysis, Amer. J. Math. Manag. Sci. 4(1984)345–373.

[18] P. Shaman and R.A. Stein, The bias of autoregressive coefficient estimators, J. Amer. Statist. Assoc. 83(1988)842–848.

[19] M. Wax, Order selection for AR models by predictive least squares, IEEE Trans. Acoustic, Speech, and Signal Proc. ASSP-36(1988)581–588.

[20] W.S. Wei, *Time Series Analysis – Univariate and Multivariate Methods* (Addison–Wesley, Redwood City, CA, 1990).

[21] M. Yuan and B.L. Nelson, Multiple comparisons with the best for steady-state simulation, ACM Trans. Mod. Comp. Simul. 3(1993)66–79.