On the Finite-Sample Accuracy of Nonparametric Resampling Algorithms for Economic Time Series

Jeremy Berkowitz
Federal Reserve Board [#]

Ionel Birgean
University of Michigan [×]

Lutz Kilian
University of Michigan [+]

January 24, 1999

ABSTRACT: In recent years, there has been increasing interest in nonparametric bootstrap inference for economic time series. Nonparametric resampling techniques help protect against overly optimistic inference in time series models of unknown structure. They are particularly useful for evaluating the fit of dynamic economic models in terms of their spectra, impulse responses, and related statistics, because they do not require a correctly specified economic model. Notwithstanding the potential advantages of nonparametric bootstrap methods, their reliability in small samples is questionable. In this paper, we provide a benchmark for the relative accuracy of several nonparametric resampling algorithms based on ARMA representations of four macroeconomic time series. For each algorithm, we evaluate the effective coverage accuracy of impulse response and spectral density bootstrap confidence intervals for standard sample sizes. We find that the autoregressive sieve approach based on the encompassing model is most accurate. However, care must be exercised in selecting the lag order of the autoregressive approximation.

[#] Jeremy Berkowitz, Stop 91, Federal Reserve Board, Washington, DC 20551. Phone: (202) 736-5581. Fax: (202) 728-5887. E-mail: m1jmb02@frb.gov.

[×] Ionel Birgean, Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220. Phone: (734) 763-6478. Fax: (734) 764-2769. E-mail: ibirgean@umich.edu.

[+] Lutz Kilian, Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220. Phone: (734) 764-2320. Fax: (734) 764-2769. E-mail: lkilian@umich.edu (address for correspondence).

1. Introduction

In recent years, there has been increasing interest in nonparametric bootstrap inference for economic time series. Nonparametric resampling techniques help protect against overly optimistic inference in time series models of unknown structure. They are particularly useful for evaluating the fit of dynamic economic models in terms of their spectra, impulse responses, and related statistics, because they do not require a correctly specified economic model (see Diebold and Kilian (1997), Diebold, Ohanian, and Berkowitz (1998)).

Notwithstanding the potential advantages of nonparametric bootstrap methods, their reliability in samples as small as those encountered in applied econometrics is questionable. Moreover, after several years of rapid advances in statistical methods, applied researchers face a dizzying array of alternative nonparametric bootstrap approaches whose accuracy in finite samples is unknown. In this paper, we explore the relative merits of several of these bootstrap procedures. The methods under consideration include the blocks-of-blocks bootstrap, the stationary bootstrap, two versions of the autoregressive (AR) sieve bootstrap (the encompassing model sieve and the AIC model sieve bootstrap), the Cholesky factor bootstrap, and three versions of a hybrid method called the pre-whitened block bootstrap. The main objective of the paper is to offer guidance to applied researchers interested in economic applications of these techniques. We also aim to provide insights for econometric theorists interested in the further development of nonparametric resampling methods.

To the best of our knowledge, this study is the first to present simulation evidence of the relative accuracy of nonparametric resampling algorithms. Our data generating processes (DGPs) are based on autoregressive-moving average (ARMA) representations of four macroeconomic time series. We view the ARMA framework as a natural benchmark for a preliminary study such as this. Presumably, no applied user would feel comfortable employing a resampling method that performs poorly for the widely studied ARMA model. Thus, the ARMA benchmark may be viewed as a necessary, but not sufficient test for the accuracy of nonparametric resampling algorithms. For each algorithm, we evaluate the effective coverage accuracy of bootstrap confidence intervals for spectral density and impulse response estimates. These statistics may be viewed as leading examples of the types of statistics of interest in applied work.

The remainder of the paper is organized as follows: In section 2 we briefly review the implementation of the nonparametric resampling algorithms to be analyzed in the Monte Carlo study. Section 3 contains a discussion of the design of the Monte Carlo study, and section 4 a summary of the simulation results. We also stress the importance of the choice of the lag order for the accuracy of the AR sieve bootstrap method. Section 5 contains the conclusion.

2. A Review of Nonparametric Resampling Methods

2.1. Autoregressive Sieve Bootstrap

Bühlmann (1997) and Bickel and Bühlmann (1995) recently adopted an idea dating back to the work of Akaike (1969) and Parzen (1974): Instead of imposing a fixed finite-dimensional ARMA model on the data, they employ the so-called "method of sieves".[1] They postulate that the data $Y = (y_1, \ldots, y_T)'$ can be described by an infinite-dimensional autoregressive process, which in turn can be approximated by a sequence of finite-dimensional models. In practice, this procedure involves fitting autoregressive processes of order p, where $p = p(T)$ is a function of the sample size T with $p(T) \rightarrow \infty$, and $p(T) = o(T)$ as $T \rightarrow \infty$. The idea of the sieve can be used to give conventional bootstrap algorithms designed for finite-order autoregressive models a nonparametric interpretation. The autoregressive sieve bootstrap procedure can be summarized as follows:

1) Determine the estimate $\hat{p}$ of the lag order p subject to the upper bound $\bar{p}_T$, and use the ordinary least-squares method to estimate the approximating autoregressive sieve model

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_{\hat{p}} y_{t-\hat{p}} + \varepsilon_t$$

where all deterministic components are assumed to have been removed.

2) Generate bootstrap innovations ε_t^* by resampling with replacement from the empirical residuals $\hat{\varepsilon}_t = y_t - \hat{a}_1 y_{t-1} - \ldots - \hat{a}_{t-\hat{p}} y_{t-\hat{p}}$

[1] Similar procedures have been proposed in Swanepoel and van Wyk (1986), Paparoditis and Streitberg (1992) and Kreiss (1992).

3) Generate a random draw for the vector of $\hat{p}$ initial observations $Y_0^* = (y_1^*, \ldots, y_{\hat{p}}^*)$.[2]

4) Conditional on Y_0^*, the estimated regression coefficients, and the sequence of bootstrap innovations, ε_t^*, generate a sequence of pseudo data of length T from the recursion:

$$y_t^* = \hat{a}_1 y_{t-1}^* + \hat{a}_2 y_{t-2}^* + \ldots + \hat{a}_{\hat{p}} y_{t-\hat{p}}^* + \varepsilon_t^*$$

For this sequence of pseudo data, calculate the statistic of interest.

5) Repeat steps 2-4 until the empirical distribution of the statistic of interest is approximated to the desired degree of accuracy.

This *autoregressive sieve bootstrap* is model-free within the class of linear MA(∞) processes with polynomial decay. Bühlmann (1997) proves that the sieve bootstrap gives correct approximations to the distributions of smooth functions of linear statistics of the data. Under the more restrictive assumption of exponential decay, Paparoditis (1996) proves that this procedure delivers an asymptotically valid bootstrap approximation to the distribution of the autoregressive coefficients and to the distribution of the implied moving average coefficients.

A key design element in the sieve bootstrap procedure is the choice of the upper bound $\overline{p}_T \geq p(T)$. Asymptotic theory sheds some light on the rate at which the lag order must increase with the sample size. For example, Paparoditis (1996) establishes that asymptotically valid bootstrap inference about reduced form impulse responses is possible provided that $\overline{p}_T = o(T^{1/7})$. Of course, such results for the rate of growth of the lag order do not help us in selecting a value for $\overline{p}_T$ for a given fixed sample size T. For the baseline results in this paper we follow the convention of allowing for up to 12 lags for the equivalent of twenty years of monthly data. We will also provide some sensitivity analysis and develop further guidelines for the choice of $\overline{p}_T$ for a given sample size.

One way of implementing the sieve bootstrap is simply to fit the AR($\overline{p}_T$) sieve at all stages of the analysis. We will refer to this version of the sieve bootstrap as the *encompassing*

---

[2] See Berkowitz and Kilian (1996) for a discussion of alternative methods for choosing Y_0^*.

model sieve bootstrap.[3] An alternative approach is to use information criteria to economize on the number of lags. For example, Bühlmann proposes selecting the lag order $p = p(T)$ of the approximating autoregressive model based on the Akaike Information Criterion (AIC) subject to the upper bound $\bar{p}_T \geq p(T)$. This *AIC model sieve bootstrap* procedure has the potential advantage of greater parsimony and efficiency compared to the encompassing model. A potential drawback of this procedure is the tendency of the AIC to underfit in small samples (see Kilian, 1997). Thus, we will employ both versions of the sieve bootstrap in section 4.

In addition, we follow a recommendation of Kilian (1998b) for bootstrapping finite lag order autoregressions and replace the least-squares estimator in step (4) by a suitably bias-corrected estimator, denoted by $\tilde{a}_i, i = 1, \ldots, \hat{p}$. The resulting bootstrap DGP is:

$$y_t^* = \tilde{a}_1 y_{t-1}^* + \tilde{a}_2 y_{t-2}^* + \ldots + \tilde{a}_{\hat{p}} y_{t-\hat{p}}^* + \varepsilon_t^* .$$

This modification is necessary in practice because the least-squares estimator of the approximating autoregressive model will tend to be biased in small samples, and therefore systematically different from the parameters of the best approximating model.

2.2. Blocks of Blocks Bootstrap

The blocks of blocks bootstrap is a nonparametric method designed to handle weakly dependent stationary observations. The properties and extensions of this bootstrap technique have been analyzed by Künsch (1989), Politis and Romano (1990, 1992a,b), Liu and Singh (1992), Politis, Romano and Lai (1992), Léger, Politis and Romano (1992), Bühlmann and Künsch (1995, 1996), and Bühlmann (1994, 1995).

Let $\{y_t\}_{t=-\infty}^{\infty}$ be a stationary stochastic process and $(y_1, y_2, \ldots, y_T)$ an observed sample of size T. We begin by describing the simplest form of the block bootstrap. The idea of the original moving blocks bootstrap (MBB) of Künsch (1989) and Liu and Singh (1992) is to construct bootstrap replicates of the data by selecting blocks of length l randomly with replacement from the original data.

[3] This approach is similar in spirit to Hjorth's (1994) idea of using the encompassing autoregressive model as the bootstrap data generating process.

1) Define $b = T - l + 1$ overlapping blocks of data $x_t = (y_t, y_{t+1}, \ldots, y_{t+l-1})$ of length l.

2) Draw with replacement from the blocks $(x_1, x_2, \ldots, x_b)$ to form a sequence of pseudo data $\{y_t^*\}_{t=1}^T = (x_1^*, x_2^*, \ldots, x_r^*)$ of length $T = rl$.

3) Calculate the statistic of interest for $\{y_t^*\}_{t=1}^T$.

4) Repeat steps 2-3 until the empirical distribution of the statistic of interest is approximated to the desired degree of accuracy.

This simple procedure works well if the statistic of interest is a symmetric function of the observations and hence does not depend on the order of the observations. A well-known example is the sample mean. Problems arise in the implementation of the MBB procedure if the statistic of interest $\theta_T = \theta_T(y_1, y_2, \ldots, y_T)$ depends on the time order of observations. Least-squares estimators of the parameters of AR models, sample autocorrelations, spectral density estimators based on autocovariance smoothing or periodogram smoothing, or on Welch's (1967) method of averaging over short periodograms, are all examples from this class of estimators. For such estimators, resampling blocks of l consecutive observations from the original sample will cause an important bias in the estimate θ_T^* of θ_T. The bias occurs because the bootstrap procedure treats consecutive blocks as conditionally independent, resulting in large jumps between blocks. These jumps in the bootstrap data bias the bootstrap estimator θ_T^*. This bias is shown to be quantitatively important in Berkowitz and Kilian (1996) even for sample sizes as large as $T = 480$.

For this reason the simplest form of the MBB is not suitable for most problems in applied time series econometrics. One solution to the bias problem was proposed by Künsch (1989), and later extended by a number of authors. Suppose the statistic of interest, θ^m, depends on some m-dimensional marginal distribution, $F^m(y_1, y_2, \ldots, y_m)$, with m fixed. Then θ^m is a symmetric function of m-dimensional blocks of consecutive observations. Therefore, as suggested by Künsch (1989), the distribution of this statistic can be approximated by using the distribution of the empirical marginal F_T^m obtained by resampling blocks of l consecutive m-tuples drawn from the set of all possible m-tuples, $\{y'_1, y'_2, \ldots, y'_{T-m+1}\}$, where $y'_t = (y_t, y_{t+1}, \ldots, y_{t+m-1})$. Specifically, let S_j be the integer part of a random variable that is iid uniformly distributed on the interval

[1, T-m+2-l]. Suppose that m = 2 for example. Then for any draw S_j, the block of l consecutive m-tuples is defined by:

$$\begin{bmatrix} y'_{S_j} \\ y'_{S_{j+1}} \\ \vdots \\ y'_{S_{j+l-1}} \end{bmatrix} = \begin{bmatrix} y_{S_j} & y_{S_{j+1}} \\ y_{S_{j+1}} & y_{S_{j+2}} \\ \vdots & \vdots \\ y_{S_{j+l-1}} & y_{S_{j+l}} \end{bmatrix}$$

This resampling plan is often referred to as the *blocks-of-blocks bootstrap*. If θ^m can be expressed as an average of some function $\phi(y_t')$, this procedure is essentially equivalent to one of resampling blocks of l consecutive sub-statistics $\phi(y_t')$, where t = T – m + 1 (see Politis, Romano and Lai, 1992).

Künsch (1989) proved the consistency of the blocks-of-blocks bootstrap for linear statistics described as functionals of m-dimensional marginals, with m fixed. Politis and Romano (1992a,b) extended the analysis to the case when m is not constant, but increases at a suitable rate with the sample size T. They proved the consistency of the blocks of blocks bootstrap for the problem of estimating the spectral density function using the Welch (1967) method. Bühlmann and Künsch (1995) establish the consistency of the blocks-of-blocks bootstrap method for statistics which estimate a parameter of the entire distribution of a stationary time series process.

A key issue in the implementation of the blocks of blocks bootstrap is the choice of the design parameter m. The appropriate block size m in general depends on the statistic of interest. For example, if the statistics of interest are the parameters of an AR(m-1) sieve process,

$$y_t = a_1^m y_{t-1} + a_2^m y_{t-2} + \ldots + a_{m-1}^m y_{t-m-1} + \varepsilon_t^m$$

the bootstrap estimators are given by:

$$\left(a_1^{m*}, a_2^{m*}, \ldots, a_{m-1}^{m*}\right) = \arg\min_{a_k} \sum_{j=1}^{b} \sum_{t=1}^{l} \left(y_{S_j+t+m-1} - \bar{y} - \sum_{k=1}^{m} a_k \left(y_{S_j+t+m-1-k} - \bar{y} \right) \right)^2,$$

where $\bar{y}$ is the sample mean, b = (T-m+1)/l is an integer, and S_j is the integer part of a random variable that is iid uniformly distributed on the interval [0, T-m-l+1]. Based on these bootstrap estimators, one can further compute impulse response estimates or spectral density estimates using the autoregressive sieve approximation. Note that by resampling m-tuples of the data we preserve the correlation between the regressand y_t and the regressors (y_{t-1}, …, y_{t-m-1}) and greatly reduce the bias problem that invalidates the naïve MBB in this case. Given the overlap between m-tuples, however, we need to choose l to be suitably large. In the absence of analytic results for the choice of l for a given sample size, we adopt the convention of setting $l = (T\text{-}m)^{1/3}$ for $T = 240$, where $T - m$ is the number of m-tuples (see Li and Maddala, 1996).

Similarly, the bootstrap estimator of the spectral density based on averaged periodograms or averaged autocovariances may be obtained by first forming m-tuples $y_t' = (y_{t,}y_{t+1},\ldots,y_{t+m-1})$, with m equal to the truncation lag, and then resampling blocks of size l out of these m-tuples. Our rules for the choice of the design parameters m and l for $T = 240$ are summarized in Table 1.

Finally, for spectral density estimators based on averages of short periodograms as proposed by Welch (1967), a modified version of the blocks of blocks bootstrap procedure has been suggested by Politis, Romano and Lai (1992). This procedure includes the following steps:

(1) Form the m-tuples $y_{i,m,L}' = (y_{(i-1)L+1,}y_{(i-1)L+2},\ldots,y_{(i-1)L+m})$, obtained from y_t by a window of width m which is moving at lags L at a time, where $i = 1, \ldots, Q$; $Q = [(T\text{-}m)/L] + 1$; and [.] denotes the integer part.

(2) Compute

$$\theta_i^{m,L}(\omega) = \frac{1}{2\pi m}\left|\sum_{t=(i-1)L+1}^{(i-1)L+m} y_t e^{-jt\omega}\right|^2,$$

for $i = 1, \ldots, Q$, and form the spectral density estimator:

$$\bar{\theta}_T(\omega) = \frac{1}{Q}\sum_{i=1}^{Q}\theta_i^{m,L}(\omega).$$

(3) Define B_j, where $j = 1, \ldots, b$ and $b = [(Q\text{-}l)/h]+1$, to be the block of l consecutive values of $\theta_i^{m,L}$, starting from $\theta_{(j-1)h+1}^{m,L}$.

(4) Sample with replacement k blocks from the set $\{B_1, B_2, \ldots, B_b\}$ and obtain the sequence of bootstrap statistics $\theta_1^{m,l*}, \theta_2^{m,l*}, \ldots, \theta_q^{m,l*}$, where $q = kl$ and $k = [Q/l]+1$. Construct

$$\bar{\theta}_q^{*}(\omega) = \frac{1}{q}\sum_{i=1}^{q}\theta_i^{m,l*}(\omega).$$

(5) The distribution of the bootstrap statistics $\sqrt{q}\,\bar{\theta}_q^* = \left(1/\sqrt{q}\right)\sum_{i=1}^{q}\theta_i^*$ approximates the sampling distribution of $\sqrt{Q}\bar{\theta}_T$.

In our Monte Carlo study, the design parameters for this procedure are chosen following the practical guidelines in Politis and Romano (1992b), and Politis, Romano and Lai (1992). Specifically, the choice of m for sample size T = 240 is based on some exploratory analysis of the spectral density estimate for different values of m. The lag L is chosen such that the segments overlap by one half of their length, i.e., $L = m/2$, a procedure suggested by Welch (1967) to obtain near maximum reduction in the variance. In our study we chose m = 40. The parameter values are summarized in Table 2.

2.3. Stationary Bootstrap

Even if the underlying population data generating process is stationary, the pseudo data generating process based on the conventional block bootstrap may not be. The problem is that the algorithm treats blocks of m-tuples as conditionally independent. Thus, the pseudo-data will in general feature a discontinuity every lm observations. This regularity is akin to introducing a spurious periodicity in the pseudo-data. The fact that pseudo-observations near the point of discontinuity will have a different joint distribution than pseudo-observations near the center of a given block means that the bootstrap DGP is not stationary. In order to address this problem, Politis and Romano (1994) introduced the *stationary bootstrap*. What distinguishes this procedure from other block resampling plans is that the block length l is stochastic. Treating l as random spreads the periodicity across frequencies and eliminates the source of the nonstationarity. In their paper, Politis and Romano proposed resampling blocks of l consecutive m-tuples of the form $y'_{t,}y'_{t+1},\ldots,y'_{t+l-1}$, for given t, where l is a random variable having the geometric distribution. The probability of the event $\{l=d\}$ is $(1-p)^{d-1}p$, for d = 1, 2, …, and a fixed number $p \in [0,1]$. The average length of the blocks, $1/p$, plays the same role as the fixed parameter l in the original blocks-of-blocks bootstrap. In our simulations we use $p = 1/l$, following a convention in the literature (see Li and Maddala (1997), Legér et al. (1992)).

2.4. Pre-Whitened Block Bootstrap

We already discussed one solution to the problem of bias that arises in linking together blocks of l consecutive observations from the original sample. That solution involved resampling blocks of blocks. In this section, we present an alternative solution based on pre-whitening. As Berkowitz and Kilian (1996) showed, the bias problem of the naïve MBB is particularly acute for economic time series that are relatively short, but highly persistent. For such series, choosing a block size l that is too small, will destroy the dependence structure of the underlying process. To see this clearly, consider the limiting case when $l = 1$. In that case, the block bootstrap reduces to iid resampling of the observations, and all dependence in the data is lost in the process of resampling. Hence, it is important to select a large enough value for l in practice. However, there is a tradeoff. The larger l, the smaller is the number of blocks available for resampling. Thus, block resampling will not provide a good estimate of the distribution of interest for l large.

One solution to this problem is known as the *pre-whitened block bootstrap* (see Davison and Hinkley (1997), p. 397). The central idea is to apply an autoregressive filter to the time series of interest *before* applying the moving blocks bootstrap. The filter removes much of the persistence in the data, leaving a sequence of residuals, $\hat{\varepsilon}_t$, that is much closer to white noise than the original data. On these residual data, the block bootstrap may be applied using relatively small block sizes l, which allows the construction of a larger number of blocks, b, without destroying the dependence structure in the data. Following the resampling of the residual blocks, the sequence of bootstrap innovations ε_t^* is applied to the estimated autoregressive filter to "re-color" the pseudo data. In practice, we employ three types of filters: the AR(1) model proposed by Davison and Hinkley (1997), the encompassing model sieve and the AIC model sieve.

The use of the AIC model sieve and the encompassing model sieve as a pre-whitening device is a natural alternative to the use of the AR(1) model. An alternative motivation for the same procedure is the view that resampling blocks of residuals in the autoregressive sieve bootstrap procedure protects against underestimation of $\hat{p}$ (or against selecting a $\bar{p}_T$ that is too low). Thus, the pre-whitened block bootstrap may be viewed simply as a version of the sieve bootstrap that is designed to be more robust against underestimation of the lag order. The only change required in the sieve bootstrap algorithm described in section 2.1. is that we draw with replacement from blocks of residuals $(\varepsilon_t, \ldots, \varepsilon_{t-l+1})$ of length l rather than from individual residuals, $\hat{\varepsilon}_t$, in forming the sequence of bootstrap innovations ε_t^*.

The optimal block size l in general depends on the persistence of the pre-whitened series and on the sample size. Ideally, one would like to select the block size according to the bootstrap grid search procedure suggested in Berkowitz and Kilian (1996). That procedure, however, is not computationally feasible within the context of the Monte Carlo analysis to be conducted in section 4. For the AR(1)-pre-whitened MBB, we experimented instead with a grid of fixed values $l \in \{30, 50, 70\}$. We chose the value of l that proved optimal in terms of the coverage accuracy of the impulse response intervals for all four DGPs. For the pre-whitened MBB based on the AIC model sieve and the encompassing model sieve much smaller block sizes are likely to suffice because most of the dependence is captured by the lag structure and the residual is close to white noise. We therefore set l according to an ad hoc rule of thumb equal to the integer part of $(T-p)^{1/3}$ where p is the lag order of the autoregressive filter (see Li and Maddala, 1996, p. 139). For example, for the encompassing model filter, this rule implies a block size of $l = 6$ for $T = 240$.

2.5. Cholesky-Factor Bootstrap

A very different nonparametric bootstrap algorithm for linear MA(∞) processes with a suitable rate of decay has been proposed by Diebold, Ohanian and Berkowitz (1998). This *Cholesky-factor bootstrap* algorithm exploits the fact that any stationary time series, $Y = (y_1, \ldots, y_T)'$, can be represented, up to second order, by a Cholesky decomposition of the series' covariance matrix and a set of uncorrelated homoskedastic residuals. In other words, there exists a representation $Y \sim (\mu, \Sigma)$ with unknown mean and variance. The mean μ can be consistently estimated by the sample mean, $\bar{y}$. In order to consistently estimate the covariance matrix Σ, the number of autocovariances being estimated must grow with (but slower than) the sample size. This can be achieved by downweighting the off-diagonal elements of $\hat{\Sigma}$ and by setting autocovariances past some truncation point equal to zero. Selecting a particular sequence of weights amounts to choosing a window and a bandwidth, B. Note that the bandwidth choice in particular will affect the performance of the Cholesky factor bootstrap. Severe downweighting induces bias, while insufficient downweighting reduces efficiency. Thus in place of the lag order selection problem for the sieve bootstrap and the block size for the block bootstrap, the Cholesky factor bootstrap requires a bandwidth choice. In this paper, we use the data-based bandwidth selection procedure of Andrews (1991) and the Newey and West (1994) lag window. We also experimented with the use of fixed bandwidth parameters for a given sample size *T*.

Based on the representation $Y \sim (\mu, \Sigma)$, Diebold, Ohanian and Berkowitz suggest the following nonparametric resampling algorithm:

1. Demean the data and construct a consistent estimate of the (*TxT*) matrix Σ.
2. Form the Cholesky decomposition $\hat{P}\,\hat{P}' = \hat{\Sigma}$.
3. Calculate the (*T x 1*) vector of standardized observations $\hat{\varepsilon} = \hat{P}^{-1}(Y - \bar{y})$.
4. Generate a sequence of scaled and demeaned bootstrap observations ε^* by drawing with replacement from the elements of $\hat{\varepsilon}$.
5. Scale the ε^* draws by $\hat{P}$ and add the sample mean:

$$Y^* = \bar{y} + \hat{P}\,\varepsilon^* \sim (\bar{y}, \hat{\Sigma}).$$

6. Construct the statistic of interest for the pseudo data Y^*.
7. Repeat steps 4-6 until the empirical distribution of the statistic of interest is approximated to the desired degree of accuracy.

3. Simulation Design

The Monte Carlo study evaluates the performance of the nonparametric resampling algorithms described in section 2 for a number of statistics of wide interest in applied time series analysis. For each method and data generating process, we construct estimates of the first 48 impulse responses and of the spectral density over various frequencies. We compare the accuracy of bootstrap inference in terms of the coverage accuracy of the associated pointwise nominal 90 percent confidence intervals.[4] While there are other statistics of interest in applied time series

[4] The effective coverage rate of a confidence interval is defined as the fraction of times that the interval covers the population value of the statistic of interest in repeated trials. Ideally, a nominal 1-α percent interval should have an effective coverage rate of 1-α percent.

analysis, these two examples can be considered representative for a wide range of similar statistics, and they are of central interest in the evaluation of macroeconomic models.[5]

3.1. Data Generating Processes

The Monte Carlo study is based on four economic time series measured at monthly frequency: the nominal interest rate (3-month Treasury-bill rate), the growth rate of seasonally adjusted industrial production, the rate of inflation of seasonally adjusted consumer prices, and the rate of change in the Yen-US Dollar exchange rate.[6] We proceed by fitting ARMA models using maximum likelihood estimation and the AIC criterion for model selection. The resulting estimates are then treated as the population values in the Monte Carlo study. The data generating processes (DGPs) are summarized in Table 3. The nominal interest rate and inflation rate are modeled as ARMA(2,4) processes with dominant roots of 0.9810 and 0.9793, respectively. Growth of industrial production is modeled as an ARMA(4,1) process with a dominant root of 0.8357; the exchange rate process is an ARMA(0,1) process in log-differences. The four DGPs are representative of the types of series most commonly used in time series econometrics. Figure 1 shows the population impulse responses and spectral densities for each DGP. Simulation results are based on 1,000 Monte Carlo trials throughout. The sample size is 240 observations, corresponding to a time span of twenty years of monthly data. Note that the design of the Monte Carlo study treats all resampling methods symmetrically, since no method encompasses the true model.

3.2. Statistics of Interest

Having generated a sequence of pseudo data, we fit an autoregressive sieve approximation to the bootstrap data, bias-correct the autoregressive coefficient estimates and construct the implied impulse response estimates. For each impulse response estimate, we construct the nominal 90 percent equal-tailed confidence interval by the percentile method of Efron and Tibshirani (1993, p. 170):

---

[5] See Christiano, Eichenbaum and Evans (1997), Cogley and Nason (1995), Diebold, Ohanian, and Berkowitz (1998), King and Watson (1996), Rotemberg and Woodford (1997), Schmitt-Grohé (1998).

[6] The data source is the DRI Economics Database. The raw data codes are PRNEW, IP, FYGM3, and EXRJAN. All simulations are carried out in MATLAB. The code is available upon request.

$$P\left[H^{*-1}(\alpha/2) < \theta < H^{*-1}(1-\alpha/2)\right] = 1-\alpha,$$

where $H^{*}(\theta) = P(\hat{\theta}^{*} \leq \theta)$ is the empirical distribution function of the bootstrap estimator $\hat{\theta}^{*}$, $H^{*-1}(p)$ denotes the *pth* quantile of the distribution of $\hat{\theta}^{*}$, and α is the nominal significance level. This percentile interval has the same asymptotic coverage error as the corresponding equal-tailed percentile-*t* interval, but tends to be much more accurate in small samples (see Kilian, 1998c). Details about the construction of this interval can be found in Kilian (1998b).

We choose the sieve approximation to calculate the impulse response functions, because it is the most natural approach to constructing an impulse response in a nonparametric setting. The approximating sieve for the impulse response calculations is based on $\overline{p}$ with the exception of the two resampling algorithms based on the AIC sieve, for which we replaced $\overline{p}$ by $\hat{p}$ in order to preserve the internal consistency of the assumptions. One concern about using the sieve approximation in the construction of the bootstrap impulse response estimator is that it inadvertently may favor the sieve method of resampling. To isolate the source of the poor performance of some resampling methods we therefore also investigated the coverage rates under the counterfactual assumption that the true ARMA model is known when we calculate the bootstrap impulse responses. There was no evidence of systematically higher coverage accuracy even in that counterfactual case; indeed, coverage accuracy often deteriorated, confirming that the resampling algorithm is driving the results rather than the sieve approximation for the bootstrap impulses.

A similar issue arises in the calculation of spectral density estimates. We therefore investigated three alternative spectral density estimators for each resampling method: the autoregressive sieve estimator (without bias corrections), an estimator based on autocovariance smoothing using the automatic truncation lag selection procedure of Bühlmann (1996), and an estimator based on averages of time blocks developed by Welch (1967). The latter estimator was included because it is used in the original blocks-of-blocks bootstrap method, as proposed by Politis and Romano (1992a, b). Each of these spectral estimators is discussed in detail in Birgean and Kilian (1998). We construct confidence intervals for each spectral density estimate using the equal-tailed pointwise percentile interval of Efron and Tibshirani (1993) as well as the pointwise percentile-*t* method of Swanepoel and van Wyk (1986, p. 138):

$$\bar{\theta}_T(\omega_j) - c\sqrt{2/q}\,\bar{\theta}_T(\omega_j) \le \theta(\omega_j) \le \bar{\theta}_T(\omega_j) + c\sqrt{2/q}\,\bar{\theta}_T(\omega_j),$$

where c is defined by

$$P\left[\frac{\left|\bar{\theta}_q^*(\omega_j) - \bar{\theta}_T(\omega_j)\right|}{\sqrt{2/q}\;\bar{\theta}_q^*(\omega_j)} \le c\right] = 1 - \alpha\,,$$

and P denotes the empirical probability based on the bootstrap distribution. α is the significance level, and G is the frequency range of interest. This interval can be shown to be equivalent to the interval used by Politis, Romano and Lai (1992).

4. Simulation Evidence

4.1. Impulse Responses

Figures 2 through 5 show the coverage rates of the confidence intervals for the first 48 impulse response response coefficients. For each bootstrap method with show two results corresponding to different degrees of parsimony. The nominal probability content of 90 percent is marked as a horizontal line in each subplot. The coverage plots suggest several interesting conclusions.

First, the AIC model sieve bootstrap tends to perform worse than the encompassing model sieve bootstrap. Raising $\bar{p}$ from 12 to 15 does little to improve the relative performance of the AIC model sieve. Thus, parsimony does not seem to be a virtue in approximating infinite order autoregressive models. This result is complementary to findings in Kilian (1997) for impulse responses in finite-dimensional autoregressive models with many lags. Kilian found that the AIC had a tendency to underfit in small samples relative to the true model.

Second, the pre-whitening method of Davison and Hinkley (1997) in many cases performs reasonably well, provided the block size l is chosen large enough. Preliminary exploration for a grid of alternative values of l suggests that the optimal value of l is near 50 for our sample size of T = 240. We show the results for a block size of l = 30 and and of l = 50. Further analysis based on the same DGPs shows that pre-whitening greatly improves the accuracy of the block bootstrap relative to the naïve MBB used in Berkowitz and Kilian (1996) and relative to the stationary bootstrap equivalent of that method. This conclusion does not appear to be sensitive to the block size l used for the naïve MBB. In particular, allowing for larger values of l in the MBB method does little to improve accuracy. Notwithstanding the striking improvements relative to the naïve MBB, the AR(1) pre-whitened MBB does not perform well for all DGPs, however. For the inflation rate DGP in particular, it is dominated by the encompassing model sieve bootstrap regardless of the choice of l. Thus, the AR(1) pre-whitened MBB cannot be recommended for applied work.

To conserve space, we do not display the corresponding results for the AIC model and encompassing model filters. In short, we find that block resampling of the pre-whitened series is not effective in robustifying the AIC model sieve bootstrap, at least for our choice of l. The coverage rates are virtually unchanged whether the model residuals are resampled as blocks or individually. Similarly, block resampling of the pre-whitened series does not improve the accuracy of the encompassing model sieve bootstrap, contrary to our conjecture in section 2.[7]

Third, the accuracy of the blocks-of-blocks bootstrap is remarkably similar to that of the encompassing model sieve bootstrap for the exchange rate and output growth DGPs, but not for the inflation and interest rate DGPs. For the latter two DGPs, the coverage accuracy of the blocks-of-blocks bootstrap method can be considerably lower. Raising m (for fixed l) in some cases strongly improves the coverage accuracy of the blocks-of-blocks bootstrap. This result mirrors similar improvements in the acuracy of the encompassing model sieve bootstrap and the pre-whitened MBB in response to increases in $\overline{p}$ and in l, respectively.

Our results suggest that the blocks of blocks bootstrap (like the pre-whitened MBB) is dominated by the encompassing model sieve bootstrap in practice. Future research will have to determine to what extent the performance of the blocks of blocks bootstrap can be further improved by raising l. However, the fact that the blocks of blocks bootstrap requires the choice of two tuning parameters rather than just one (as in the case of the pre-whitened bootstrap and encompassing model sieve bootstrap), is an obvious disadvantage in practice.

Fourth, our results shed light on the importance of imposing stationarity in generating bootstrap data using the blocks-of-blocks algorithm. Recently, there has been some concern about the fact that the blocks-of-blocks bootstrap does not impose stationarity in generating bootstrap data. The stationary bootstrap was explicitly designed to overcome this drawback of

[7] Our findings do not mean that there is no role for the block resampling of residuals in autoregressive models. Even serially uncorrelated innovations may be dependent via higher moments (possibly as a result of ARCH). An interesting topic for future research would be the question of whether block resampling of the residuals may help account for such higher-order dependence in the innovations.

the blocks-of-blocks bootstrap. As Figures 2 through 5 show, imposing stationarity in generating the bootstrap data does not improve accuracy, even for highly persistent DGPs, and may actually lower it slightly. We conjecture that this result is driven by occasional draws of very low block sizes, l. The use of unreasonably short blocks is likely to cause severe small-sample bias in the estimator of interest for the reasons discussed in section 2.

Fifth, the Cholesky factor bootstrap based on Andrews' rule appears to fail dramatically, especially at higher horizons. This pattern is suggestive of underfitting (see Kilian, 1997). We therefore also experimented with fixed truncation lags. Figures 2-5 show results for $B = 15$. For the inflation rate and output growth DGP the estimated bandwidth typically is smaller than 15. As expected, coverage rates improve substantially for $B = 15$ (although not nearly enough). For the interest rate DGP the pattern is reversed. Andrews' rule typically selects a bandwidth in excess of 15 for this DGP, and coverage drops if a fixed bandwidth $B = 15$ is used instead. In fact, even for $B = 50$, the Cholesky factor bootstrap intervals for the interest rate DGP have coverage rates that are much too low. We conclude that the Cholesky factor bootstrap method, as it is commonly implemented based on Andrews' rule, is not a reliable method of constructing confidence intervals. It would be useful to investigate other means of selecting the truncation lag for this resampling procedure. Clearly, the accuracy of the intervals can be improved, perhaps substantially, by allowing for larger bandwidths. Computational constraints prohibited further systematic exploration of this issue for this paper.

Sixth, as judged by the relative performance across the four DGPs, only resampling algorithms based on the encompassing model can be recommended. The blocks of blocks bootstrap, the stationary blocks of blocks bootstrap, and the AR(1) pre-whitened MBB produce mixed results. Of the eight methods included in the study, the AIC model sieve bootstrap and the Cholesky factor bootstrap rank last. This ranking does not mean that the encompassing model sieve bootstrap method can be recommended without reservations. It is essential for the success of any method, including the encompassing model sieve method, that the approximating model be generously parameterized. For our DGPs a value of $\bar{p} = 15$ works well. We will return to this point in section 4.3.

Moreover, even the best method can be counted on to deliver accurate confidence intervals only for the first 17 months. For higher horizons, coverage accuracy may be slightly too low or in many cases excessively high. Notably, for the exchange rate DGP the bootstrap intervals contain the true value virtually all the time at long horizons. This fact is less worrisome from an applied user's point of view than it may seem. 100 percent coverage usually is suggestive of a very inefficient and wide interval. In this case, however, the bootstrap confidence intervals tend to be very close to zero and narrow, despite the excess coverage, and the impulse responses would be considered economically insignificant in any case. Overall, Figures 2 through 5 provide compelling evidence that nonparametric bootstrap inference for impulse responses can be remarkably accurate even in fairly small samples if we allow for enough lags

4.2. Spectral Densities

Note that for each of the seven resampling method there are two confidence intervals (percentile and percentile-t) that can be combined with each of the three spectral density estimators (the autoregressive sieve method, the autocovariance-based estimator based on Bühlmann's (1996) automatic truncation lag selection procedure, and the Welch (1967) method) and four DGPs, resulting in a total of 168 different combinations of DGP, confidence interval, spectral density estimator, and resampling method for a given sample size. The large number of results makes it impractical to include all of them in the paper. Instead, we summarize some of the main findings and focus on the most promising results.

A key finding is that neither the Welch estimator nor Bühlmann's estimator of the spectral density can be recommended. Use of these estimators tends to result in extremely poor coverage rates for at least one DGP, regardless of the resampling method and confidence interval used. In contrast, the autoregressive sieve estimator of the spectral density results in very accurate coverage for some resampling methods, regardless of the DGP. This finding is consistent with a recent study of spectral density estimators by Birgean and Kilian (1998) who show that the autoregressive sieve approach produces the most accurate point estimates of spectral densities at frequencies other than zero. We also find that the blocks of blocks bootstrap performs best in combination with the percentile interval for the sieve spectral density estimator. This particular version of the blocks of blocks bootstrap is considerably more accurate in our simulations than the version originally proposed by Politis and Romano (1992a, b) based on the percentile-*t* interval and the Welch estimator.

Figures 6 through 9 show the coverage rates of the percentile and percentile-*t* intervals for the sieve spectral density estimator for $T = 240$. To conserve space, we only show results for the less parsimonious of the two specifications considered for each bootstrap method in Figures 2 through 5. In the case of the Cholesky factor method the results shown are for $B = 15$. Using Andrews' rule to select B does not change the qualitative results.

We find that for the resampling methods that do perform well, as a rule, the percentile interval is more accurate (often by a wide margin) than the percentile-*t* interval. In fact, not a single resampling algorithm produces accurate results across all DGPs when combined with the percentile-*t* interval. In some cases, the effective coverage rate of the nominal 90 percent percentile-*t* interval is close to 0 percent, and coverage rates often are erratic. This erratic behavior is consistent with other recent work on the percentile-*t* interval (see Kilian, 1998c).

For the percentile interval, the three most accurate resampling algorithms are the encompassing model sieve bootstrap, followed by the blocks-of-blocks and the stationary bootstrap. There is little difference between these methods. The AR(1) pre-whitened MBB is somewhat less accurate than the blocks-of-blocks bootstrap. Generally, the coverage rates of the percentile interval are very close to the nominal 90 percent coverage. The only shortcoming of the percentile interval is a lack of coverage at very long and to a lesser extent at very short horizons for some DGPs. Considered jointly with the results for the impulse response confidence intervals, the results in Figures 6-9 make the encompassing model sieve bootstrap the unambiguous winner of the competition for the most reliable nonparametric resampling method.

4.3. On the Importance of the Autoregressive Lag Order for the Sieve Bootstrap

Our results suggest that the accuracy of bootstrap inference is more susceptible to the shape of the underlying impulse response function than to the dominant root of the population process. In fact, the coverage accuracy of many resampling methods is highest for the extremely persistent interest rate DGP, much lower for the equally persistent inflation DGP, and still lower for the much less persistent output growth DGP. The key factor in driving coverage accuracy appears to be the degree of parsimony imposed. For example, for the growth of industrial production series, there is a serious loss of coverage for horizons 14-28 for $\bar{p}_T = 12$. In some cases, coverage rates drop close to 55 percent. The addition of three lags dramatically improves coverage accuracy. The lowest coverage rate rises from 55 percent to 83 percent for $\bar{p}_T = 15$, all but eliminating the trough in the coverage rates that arises for $\bar{p}_T = 12$. For the inflation rate DGP coverage rates also improve by several percentage points at long horizons. Even for the exchange rate DGP there are some slight improvements. Only for the interest rate DGP we observe some minor deterioration of the coverage accuracy at short horizons.

The sieve spectral density estimator shows a similar improvement, as $\bar{p}_T$ is increased from 12 to 15. We focus on the percentile interval only, given the erratic behavior of the percentile-*t* interval. For most frequencies, the coverage rates of the percentile interval are rather close to nominal coverage even for $\bar{p}_T = 12$. Thus, the scope for improvements is limited. Nevertheless, the coverage accuracy of the percentile interval tends to further improve for $\bar{p}_T = 15$, notably at very low and very high frequencies with gains in excess of 0.10 for some frequencies.

Interestingly, the increase in average interval length that is associated with the increase in the lag order is typically small both for spectral analysis and for impulse response analysis. Thus, the common view that increasing the autoregressive lag order is bound to result in excessively wide and imprecise intervals is not supported by the simulation results in this paper.

The central message of this experiment is that the upper bounds $\bar{p}_T$ required for accurate inference about economic time series may be considerably larger than commonly assumed. We showed that with a suitably large upper bound, reliable nonparametric inference appears feasible even for the small sample sizes available in practice. The importance of the use of less parsimonious approximations is not limited to the sieve bootstrap, however. We conjecture that the accuracy of methods like the blocks-of-blocks bootstrap and the Cholesky factor bootstrap could be improved as well. In particular, the choice of *l* for the blocks-of-blocks bootstrap for a given sample size deserves further study. Given the computational requirements of such an undertaking, this topic is the appropriate subject of another study.

A key insight that distinguishes nonparametric inference about time series processes from parametric analysis is the fact that the degree of parsimony of the approximating model must decrease, as the sample size is increased. Existing theoretical results shed some light on the rate at which the lag order must increase with the sample size. For example, Inoue and Kilian (1999) establish that asymptotically valid bootstrap inference about smooth functions of VAR slope parameters and innovation variances is possible provided that $\bar{p}_T = o(T^{1/7})$. Additional experiments for T = 480 based on the DGPs of section 3 cast doubt on the practical usefulness of this result, however. We find that in some cases the lag order must increase considerably faster than allowed by the asymptotic theory in order for the coverage rates of the confidence intervals not to deteriorate relative to the results for T = 240.

5. Concluding Remarks

The main objective of this paper was to offer guidance to applied researchers interested in economic applications of nonparametric bootstrap techniques. For this purpose, we evaluated the coverage accuracy of bootstrap confidence intervals based on eight alternative nonparametric resampling algorithms. Using ARMA representations of four macroeconomic time series as our benchmark, we found that the autoregressive sieve approach based on the encompassing autoregressive model is most accurate both for impulse response and for spectral analysis. Provided care is exercised in selecting the lag order of the autoregressive approximation, this sieve method was shown to produce remarkably accurate bootstrap confidence intervals in small samples. In contrast, the Cholesky factor bootstrap was shown not to be a viable alternative method in its present form. The AIC model sieve, the blocks-of-blocks bootstrap and the stationary bootstrap performed better, but still tended to be less accurate than the sieve bootstrap based on the encompassing model. In addition to these methods, we evaluated several pre-whitened moving block bootstrap (MBB) algorithms designed to overcome the limitations of the naïve MBB in dealing with highly persistent time series of short length. This method, while appealing in principle, did not have higher accuracy than more conventional sieve bootstrap methods.

Our study suggests several conclusions worth emphasizing. First, the pre-whitened MBB and the blocks-of-blocks bootstrap represent a considerable improvement over the naïve MBB method discussed in Efron and Tibshirani (1993, p. 99) and Berkowitz and Kilian (1996), and are likely to replace the latter in most time series applications. Second, imposing stationarity in resampling blocks-of-blocks has little effect on coverage accuracy. In fact, the stationary bootstrap tends to be slightly less accurate than the conventional blocks-of-blocks bootstrap. Third, the strategy of selecting parsimonious sieve models using the AIC or related lag order selection criteria as suggested by Bühlmann (1997) does not appear promising in infinite lag-order models. Even the AIC, which is known to be the least parsimonious criterion in its class, has a tendency to underfit the sieve. Sieve models with many more lags than would ordinarily be considered appear to be required for robust bootstrap inference. A similar insight applies to the use of Andrews' rule in selecting truncation lags for the Cholesky factor bootstrap. Fourth, we compiled additional evidence that for nonlinear statistics the percentile-*t* interval tends to be less accurate in small sample than the percentile interval. Our results for the coverage accuracy of the percentile-*t* interval for spectral density estimators complement similar findings for impulse responses in Kilian (1998c). Fifth, preliminary experiments suggested that the usual asymptotic results about the rate of decrease in the parsimony of the approximating model with an increase in the sample size may perform poorly in practice. Often, faster increases in model size are needed to prevent a deterioration in coverage accuracy.

The results in this paper suggest that there is reason to be cautiously optimistic about the future of nonparametric bootstrap inference for economic time series. Our results are intended to provide a benchmark for further discussion. As more evidence for other processes accumulates and statistical methods are improved, one would undoubtedly expect our views on the suitability of these methods to evolve. More attention needs to be devoted to the choice of lag orders and other tuning parameters for a given sample size. In particular, it would be desirable to develop at least some rules of thumb for selecting an approximating model for a given sample size. For example, our simulation results tentatively suggest that, for a sample size of 240 monthly observations, 15 autoregressive lags are likely to provide a good sieve approximation for a wide range of univariate processes. In the absence of theoretical results, rules such as these will be needed to enable practitioners to apply nonparametric resampling methods in their research.

Acknowledgments

We thank Jonathan Wright and Tao Zha for helpful discussions. The views expressed in this paper do not necessarily represent those of the Federal Reserve Board or its staff.

References

Akaike, H. (1969), "Power Spectrum Estimation through Autoregressive Model Fitting," Annals of the Institute of Statistical Mathematics, 21, 407-419.

Andrews, D.W.K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," Econometrica, 59, 817-858.

Berkowitz, J., and L. Kilian (1996), "Recent Developments in Bootstrapping Time Series," Finance and Economics Discussion Series, Working Paper No. 1996-45, Board of Governors of the Federal Reserve, forthcoming: Econometric Reviews.

Bickel, P.J., and P. Bühlmann (1995), "Mixing Property and Functional Central Limit Theorems for a Sieve Bootstrap in Time Series," Technical Report 440, Department of Statistics, University of California at Berkeley.

Birgean, I., and L. Kilian (1998), "Data-Driven Nonparametric Spectral Density Estimators for Economic Time Series: A Monte Carlo Study," mimeo, Department of Economics, University of Michigan.

Bühlmann, P. (1994), "Blockwise Bootstrapped Empirical Process for Stationary Sequences", Annals of Statistics, 22, 995-1012.

Bühlmann, P. (1995), "The Blockwise Bootstrap for General Empirical Processes of Stationary Sequences", Stochastic Processes and their Applications, 58, 247-265.

Bühlmann, P. (1996), "Locally Adaptive Lag-Window Spectral Estimation," Journal of Time Series Analysis, 17, 247-270.

Bühlmann, P. (1997), "Sieve Bootstrap for Time Series," Bernoulli, 3, 123-148.

Bühlmann, P. and Künsch, H.R. (1995), "The Blockwise Bootstrap for General Parameters of a Stationary Time Series", Scandinavian Journal of Statistics, 22, 35-54.

Bühlmann, P. and H.R. Künsch (1996), "Block Length Selection in the Bootstrap for Time Series", mimeo, Seminar für Statistik, ETH Zürich.

Christiano, L.J., M. Eichenbaum and C. Evans (1997), "Modeling Money," mimeo, Department of Economics, Northwestern University.

Cogley, T., and J.M. Nason (1995), "Output Dynamics in Real-Business Cycle Models," American Economic Review, 85, 492-511.

Davison, A.C., and D.V. Hinkley (1997), Bootstrap Methods and Their Application, Cambridge University Press.

Diebold, F.X., and L. Kilian (1997), "Measuring Predictability: Theory and Macroeconomic Applications," NBER Technical Working Paper 213.

Diebold, F.X., L.E. Ohanian, and J. Berkowitz (1998), "Dynamic Equilibrium Economies: A Framework for Comparing Models and Data," Review of Economic Studies, 65, 433-451.

Efron, B., and R.J. Tibshirani (1993), An Introduction to the Bootstrap, New York: Chapman and Hall.

Hjorth, J.S.U. (1994), Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap, London: Chapman and Hall.

Inoue, A., and L. Kilian (1999), "Bootstrapping Smooth Functions of Slope Parameters and Innovation Variances in VAR(∞) Models," mimeo, Department of Economics, University of Michigan.

Kilian, L. (1997), "Impulse Response Analysis in VectorAutoregressions with Unknown Lag Order," mimeo, Department of Economics, University of Michigan.

Kilian, L. (1998a), "Accounting for Lag Order Uncertainty in Autoregressions: The Endogenous Lag Order Bootstrap Algorithm", Journal of Time Series Analysis, 19, 531-548.

Kilian, L. (1998b) "Small-Sample Confidence Intervals for Impulse Response Functions," Review of Economics and Statistics, 80, 218-230.

Kilian, L. (1998c), "Pitfalls in Constructing Bootstrap Confidence Intervals for Asymptotically Pivotal Statistics", mimeo, Department of Economics, University of Michigan.

King, R.G., and M.W. Watson (1996), "Money, Prices, Interest Rates and the Business Cycle", Review of Economics and Statistics, 78, 35-53.

Kreiss, J.P. (1992), "Bootstrap Procedures for AR(∞)-Processes," in Bootstrapping and Related Techniques, Proceedings, Lecture Notes in Economics and Mathematical Systems, eds. K.H. Jöckel, G. Rothe, and W. Sendler. New York: Springer-Verlag.

Künsch, H.R. (1989), "The Jackknife and the Bootstrap for General Stationary Observations," Annals of Statistics, 17, 1217-1241.

Léger, C., Politis, D.N. and Romano, J.R. (1992), "Bootstrap Technology and Applications", Technometrics, 34, 378-398

Li, H., and G.S. Maddala (1996), "Bootstrapping Time Series Models," Econometric Reviews, 15, 115-158.

Li, H., and G.S. Maddala (1997), "Bootstrapping Cointegrating Regressions," Journal of Econometrics, 80, 297-318.

Liu, R.Y., and K. Singh (1992), "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence," in Exploring the Limits of Bootstrap, R. LePage and L. Billard, eds. New York: John Wiley and Sons.

Newey, W.K., and K.D. West (1994), "Automatic Lag Selection in Covariance Matrix Estimation," Review of Economic Studies, 61, 631-653.

Paparoditis, E. (1996), "Bootstrapping Autoregressive and Moving Average Parameter Estimates of Infinite Order Vector Autoregressive Processes," Journal of Multivariate Analysis, 57, 277-296.

Paparoditis, E., and B. Streitberg (1992), "Order Identification in Stationary Autoregressive Moving Average Models: Vector Autocorrelations and the Bootstrap," Journal of Time Series Analysis, 13, 415-435.

Parzen, E. (1974), "Some Recent Advances in Time Series Modeling," IEEE Transactions on Automatic Control, AC-19, 723-729.

Politis, D.N., and J.P. Romano (1992a), "A General Resampling Scheme for Triangular Arrays of α-Mixing Random Variables with Application to the Problem of Spectral Density Estimation," Annals of Statistics, 20, 1985-2007.

Politis, D.N., and J.P. Romano (1992b), "A Nonparametric Resampling Procedure for Multivariate Confidence Regions in Time Series Analysis," in Computing Science and Statistics, Proceedings of the 22nd Symposium on the Interface, eds. C. Page and R. LePage, New York: Springer-Verlag.

Politis, D.N., and J.P. Romano (1994), "The Stationary Bootstrap," Journal of the American Statistical Association, 89, 1303-1313.

Politis, D.N., J.P. Romano and T. Lai (1992), "Bootstrap Confidence Intervals for Spectra and Cross-Spectra," IEEE Transactions on Signal Processing, 40, 1206-1215.

Rotemberg, J., and M. Woodford (1997), "An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy," paper presented at the Twelfth Annual Macroeconomic Conference, NBER, April 4-5, 1997.

Schmitt-Grohé, S. (1998), "The International Transmission of Economic Fluctuations: Effects of U.S. Business Cycles on the Canadian Economy," Journal of International Economics, 44, 257-287.

Swanepoel, J.W.H., and J.W.K. van Wyk (1986), "The Bootstrap Applied to Power Spectral Density Function Estimation," Biometrika, 73, 135-141.

Welch, P.D. (1967), "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short Modified Periodograms," IEEE Transactions on Audio and Electroacoustics, AU-15(2), 70-73.

Table 1: Blocks-of-Blocks Bootstrap Design Parameters

Block Sizes	Autoregressive Sieve Impulse Response Estimator	Autoregressive Sieve Spectral Estimator	Autocovariance-based Spectral Estimator
m	$\bar{p}$	$\bar{p}$	*Truncation Lag*
l	$(T\text{-}m)^{1/3}$	$(T\text{-}m)^{1/3}$	$(T\text{-}m)^{1/3}$

NOTES: m refers to the first-stage block size for the blocks-of-blocks bootstrap; l refers to the block size used in resampling the m-tuples constructed in the first stage. *Truncation lag* refers to the lag at which the weighted sum of autocovariances is truncated in order to ensure the consistency of the spectral density estimator.

Table 2: Design Parameters for Blocks-of-Blocks Bootstrap for

Welch Spectral Density Estimator

m	40
L	20
Q	11
l	3
$k = [Q/l] + 1$	4
$q = kl$	12

NOTES: Adapted from Politis, Romano, and Lai (1992). See text for a definition of the design parameters.

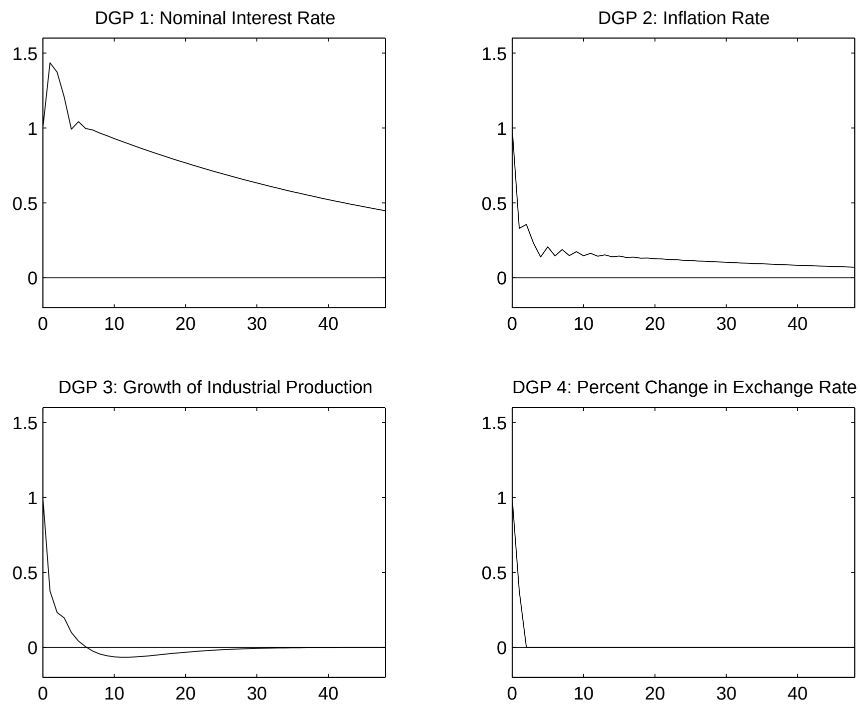
Table 3: Monthly ARMA data generating processes

Process	a_0	a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4	σ^2
Nominal Interest Rate ARMA(2,4)	0.1572	0.6197	0.3544			0.8155	0.1288	-0.1530	-0.2422	0.1967
Growth of Industrial Production ARMA(4,1)	0.0117	1.3272	-0.2668	-0.0119	-0.0945	-0.9506				0.9357
Inflation ARMA(2,4)	0.1672	0.1724	0.7901			0.1583	-0.4902	-0.0912	-0.1812	6.4564
Percent Change in Yen-Dollar Rate ARMA(0,1)	0.1437					0.3772				6.8593

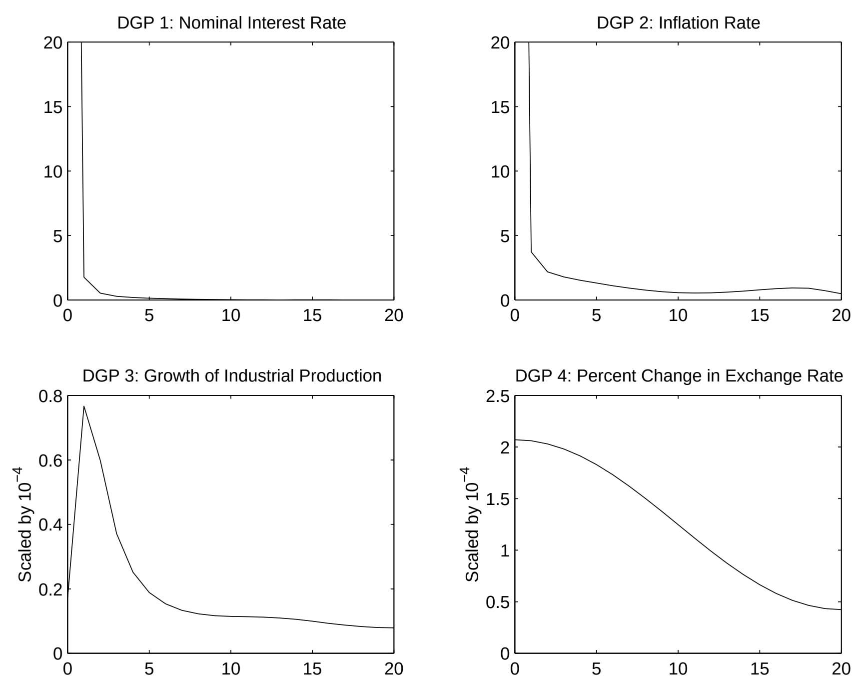
NOTES: Coefficients in the table correspond to the DGP: $y_t = a_0 + a_1 y_{t-1} + \ldots + a_p y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \ldots$
$+ b_q \varepsilon_{t-q}$. Lag orders are selected by the AIC. All data are taken from the DRI database.

Figure 1 – Data Generating Processes

a) Population Impulse Responses



b) Population Spectral Densities

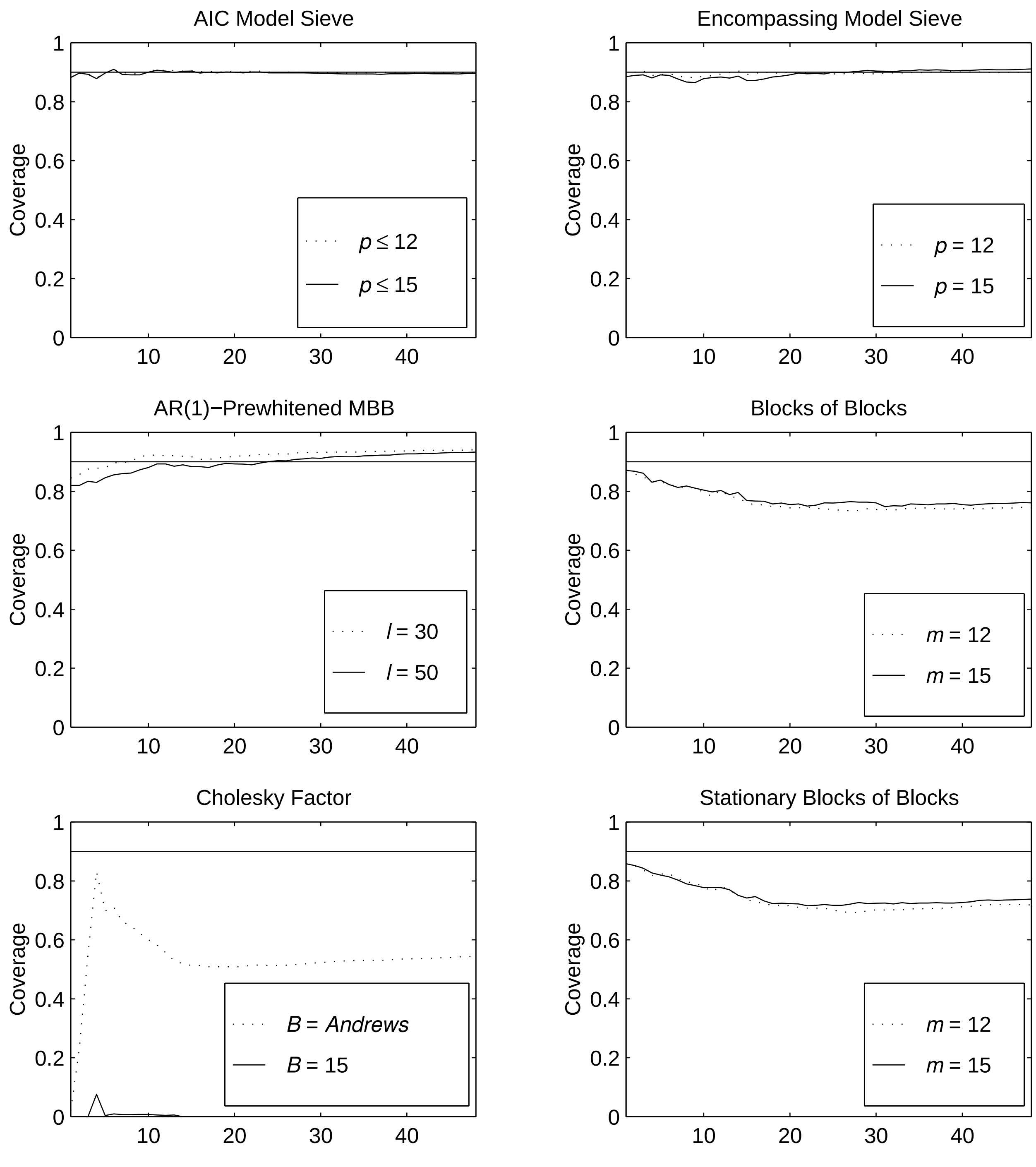


NOTES: All plots are based on the ARMA data generating processes described in Table 3. The spectral densities are defined as $f(\omega_j)$ where $\omega_j = \pi j/20$, $j = 1, \ldots, 20$.

Figure 2 – Impulse Responses

Effective Coverage Rates of Pointwise Nominal 90 % Confidence Intervals for T=240

DGP 1: Nominal Interest Rate

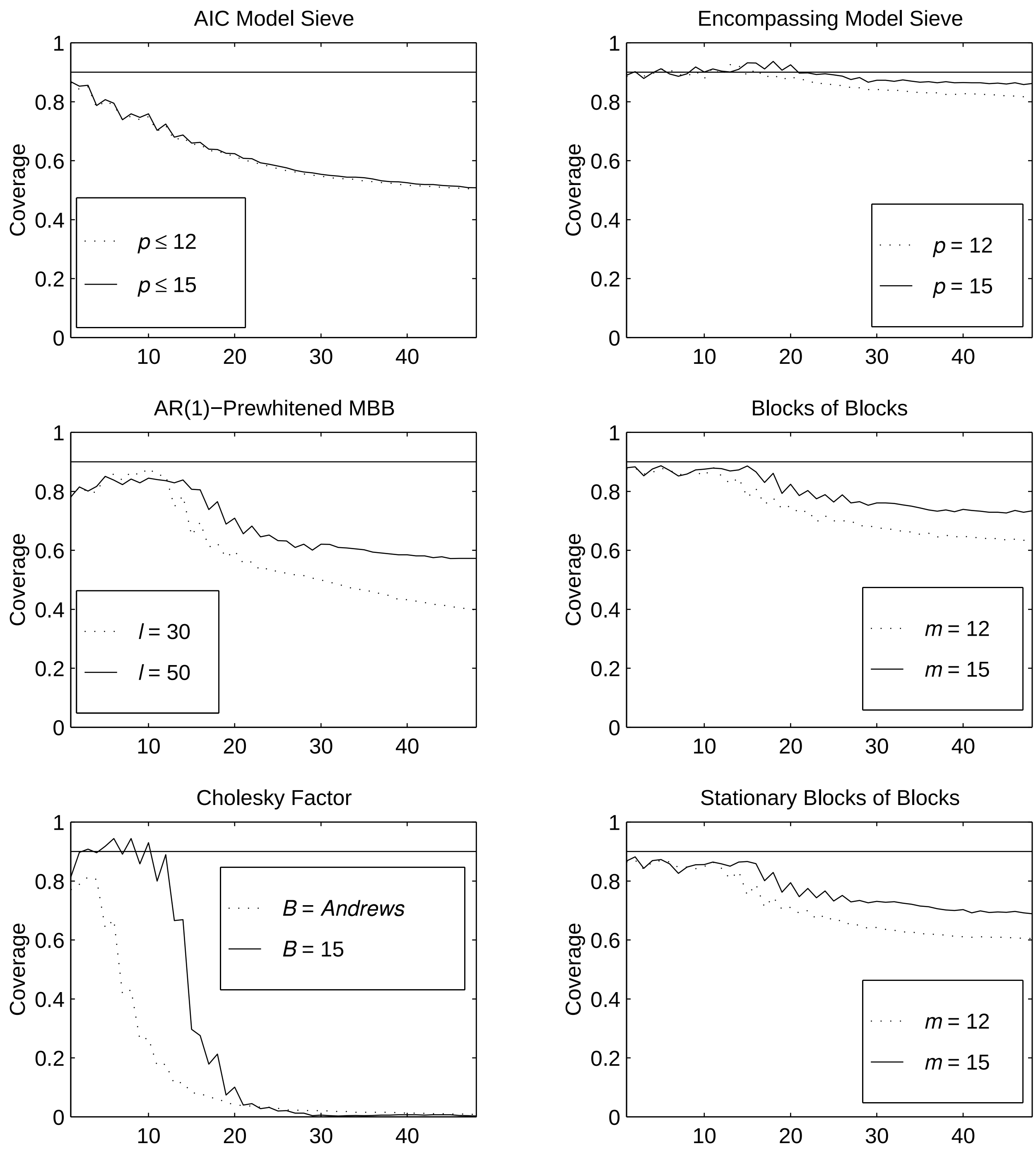


NOTES: The ARMA data generating processes are described in Table 3. The number of Monte Carlo trials is 1,000. All intervals are computed based on 1,000 bootstrap replications.

Figure 3 – Impulse Responses

Effective Coverage Rates of Pointwise Nominal 90 % Confidence Intervals for T=240

DGP 2: Inflation Rate

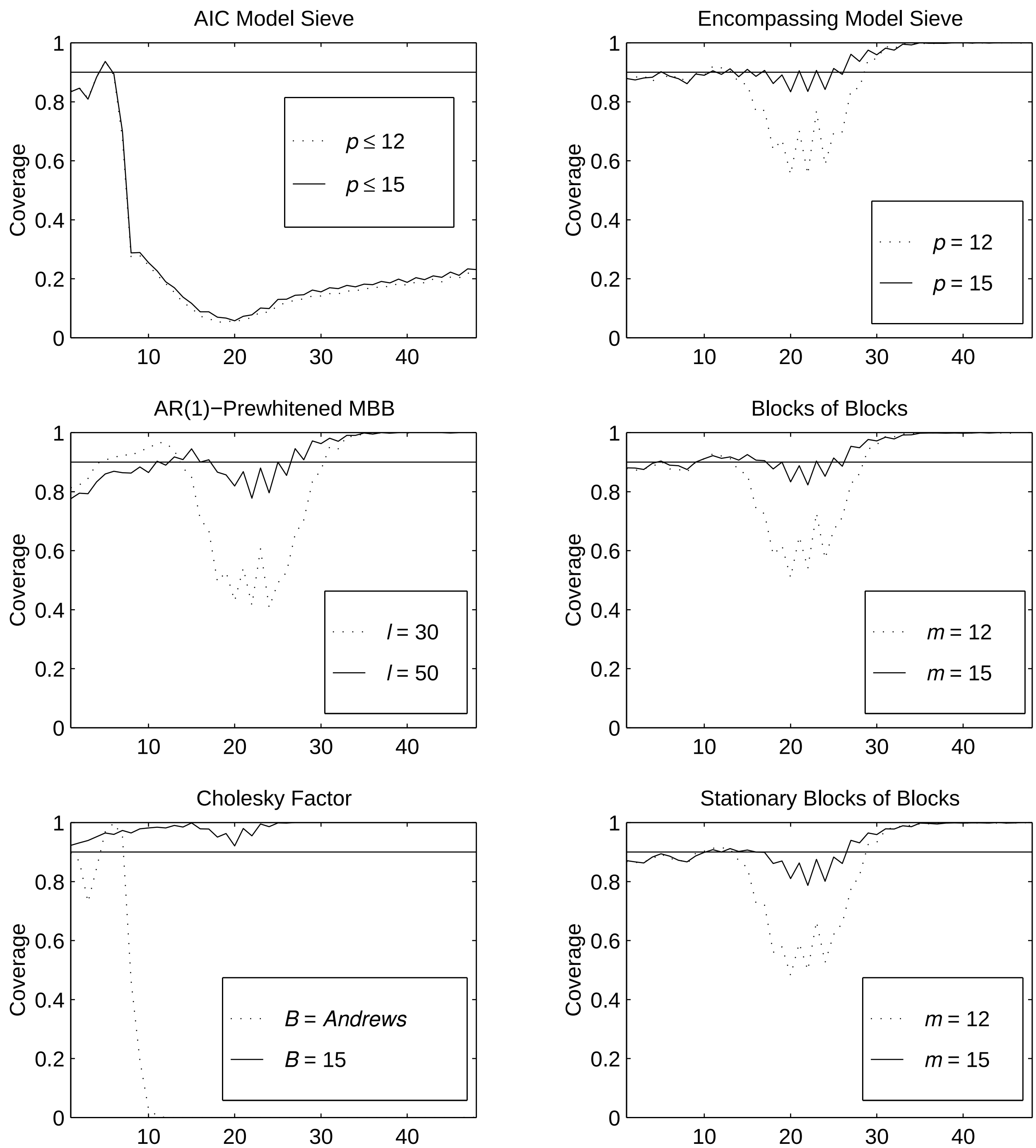


NOTES: See Figure 2.

Figure 4 – Impulse Responses

Effective Coverage Rates of Pointwise Nominal 90 % Confidence Intervals for T=240

DGP 3: Growth of Industrial Production

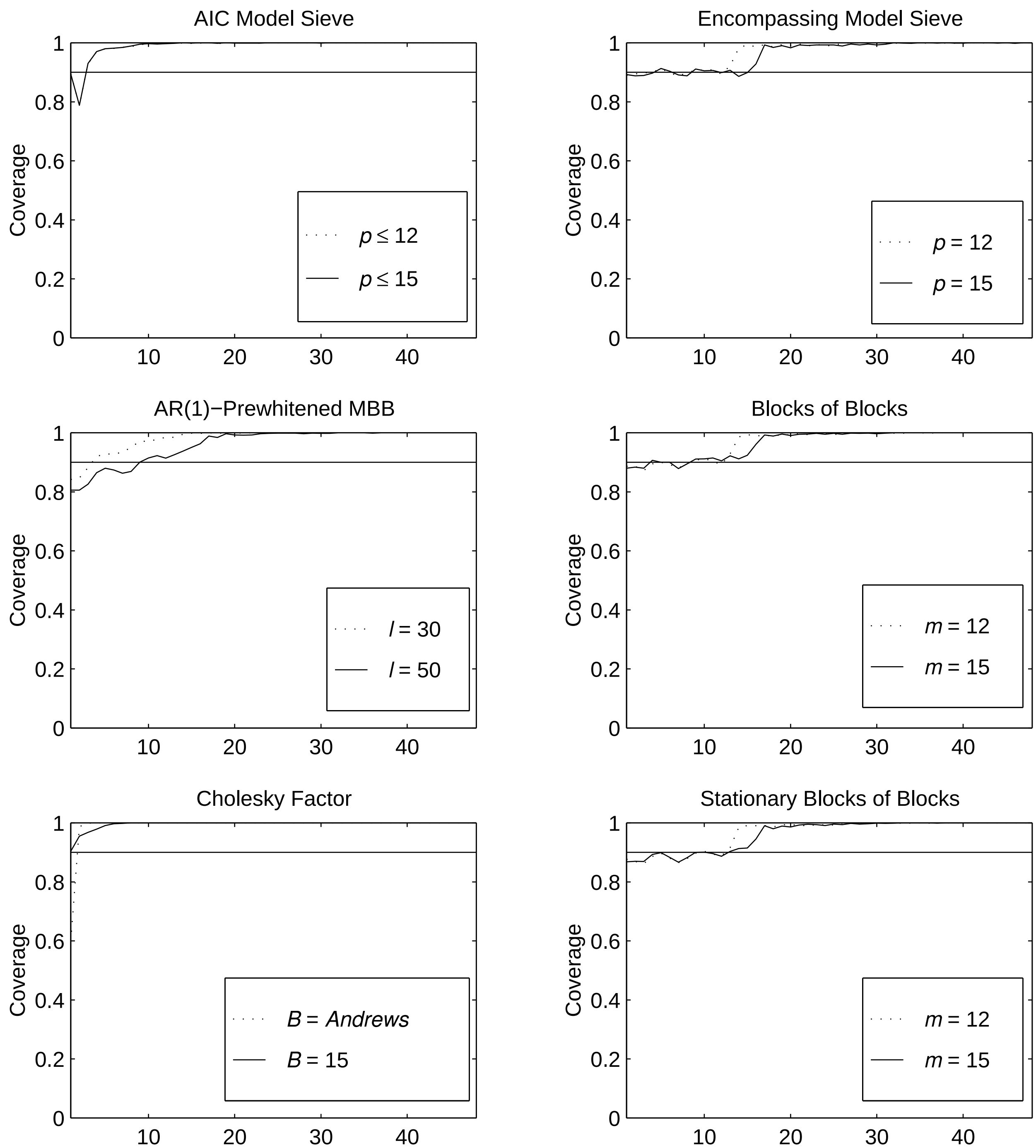


NOTES: See Figure 2.

Figure 5 – Impulse Responses

Effective Coverage Rates of Pointwise Nominal 90 % Confidence Intervals for T=240

DGP 4: Percent Change in Yen-Dollar Exchange Rate

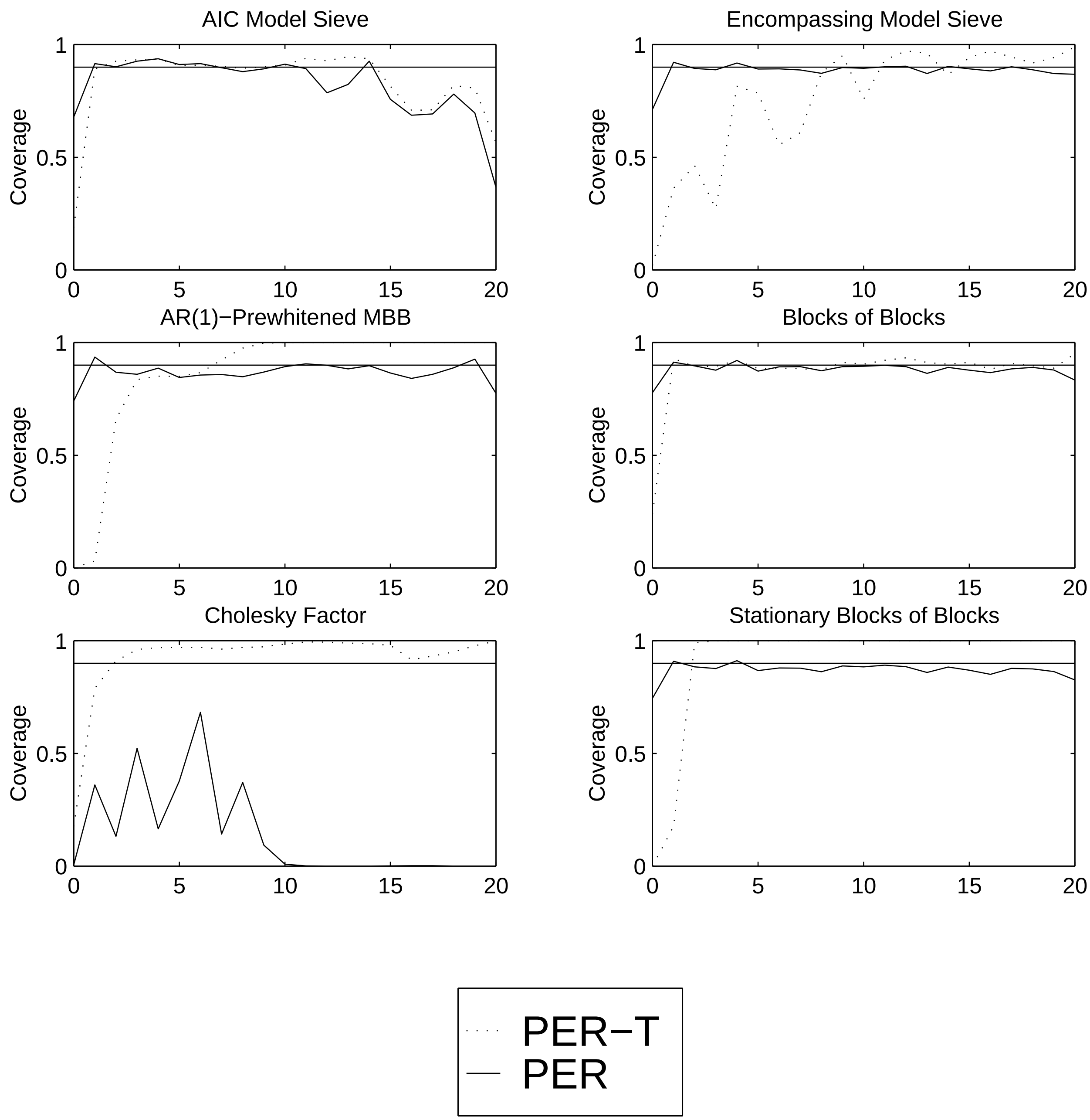


NOTES: See Figure 2.

Figure 6 – Spectral Density

Effective Coverage Rates of Pointwise Nominal 90 % Confidence Intervals for T=240

DGP 1: Nominal Interest Rate

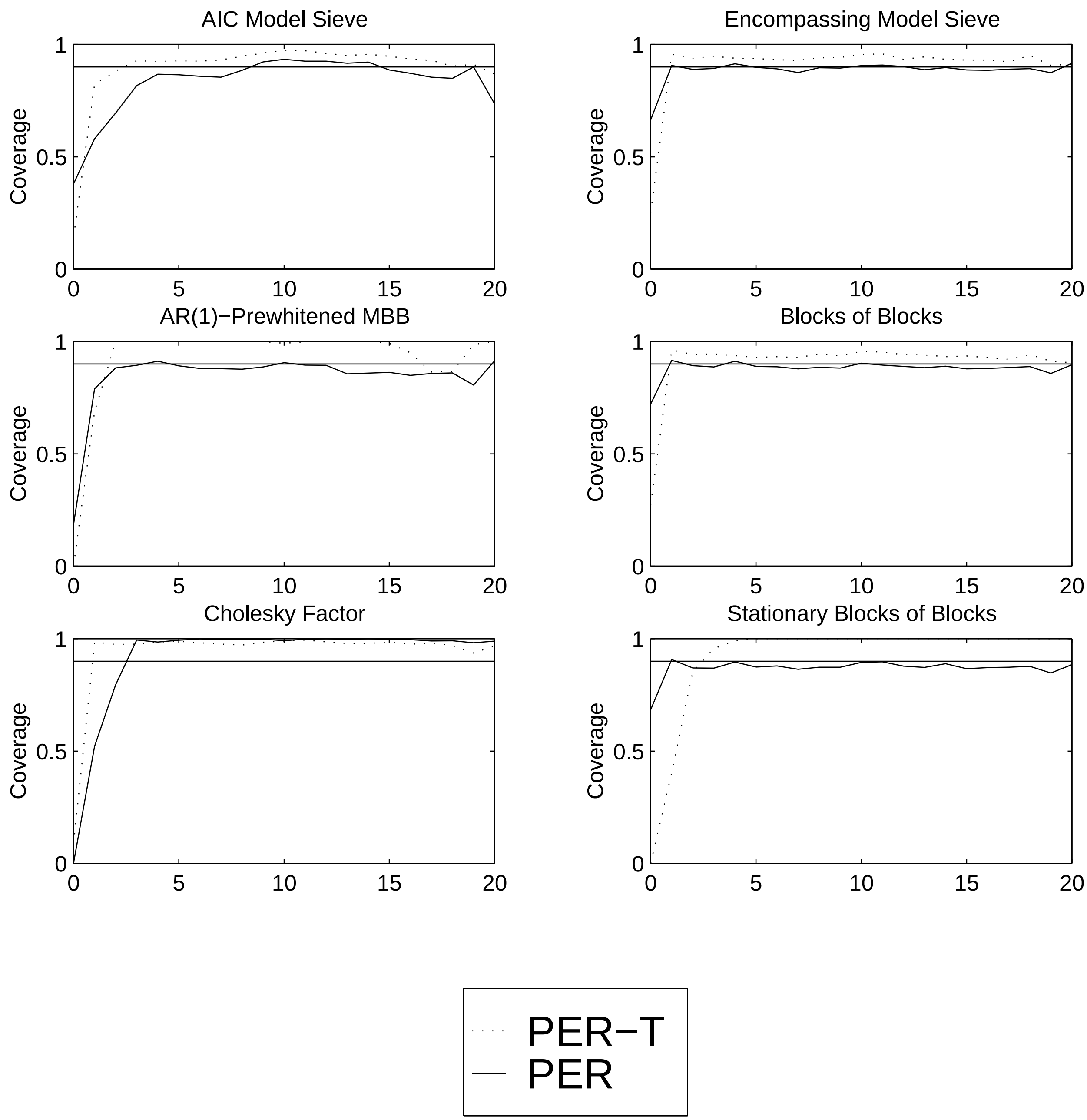


NOTES: The ARMA data generating processes are described in Table 3. The number of Monte Carlo trials is 1,000. All intervals are computed based on 1,000 bootstrap replications. PER refers to the percentile interval and PER-T to the percentile-t interval. The spectral densities are defined as $f(\omega_j)$ where $\omega_j = \pi j/20$, $j = 1, \ldots, 20$.

Figure 7 – Spectral Density

Effective Coverage Rates of Pointwise Nominal 90 % Confidence Intervals for T=240

DGP 2: Inflation Rate

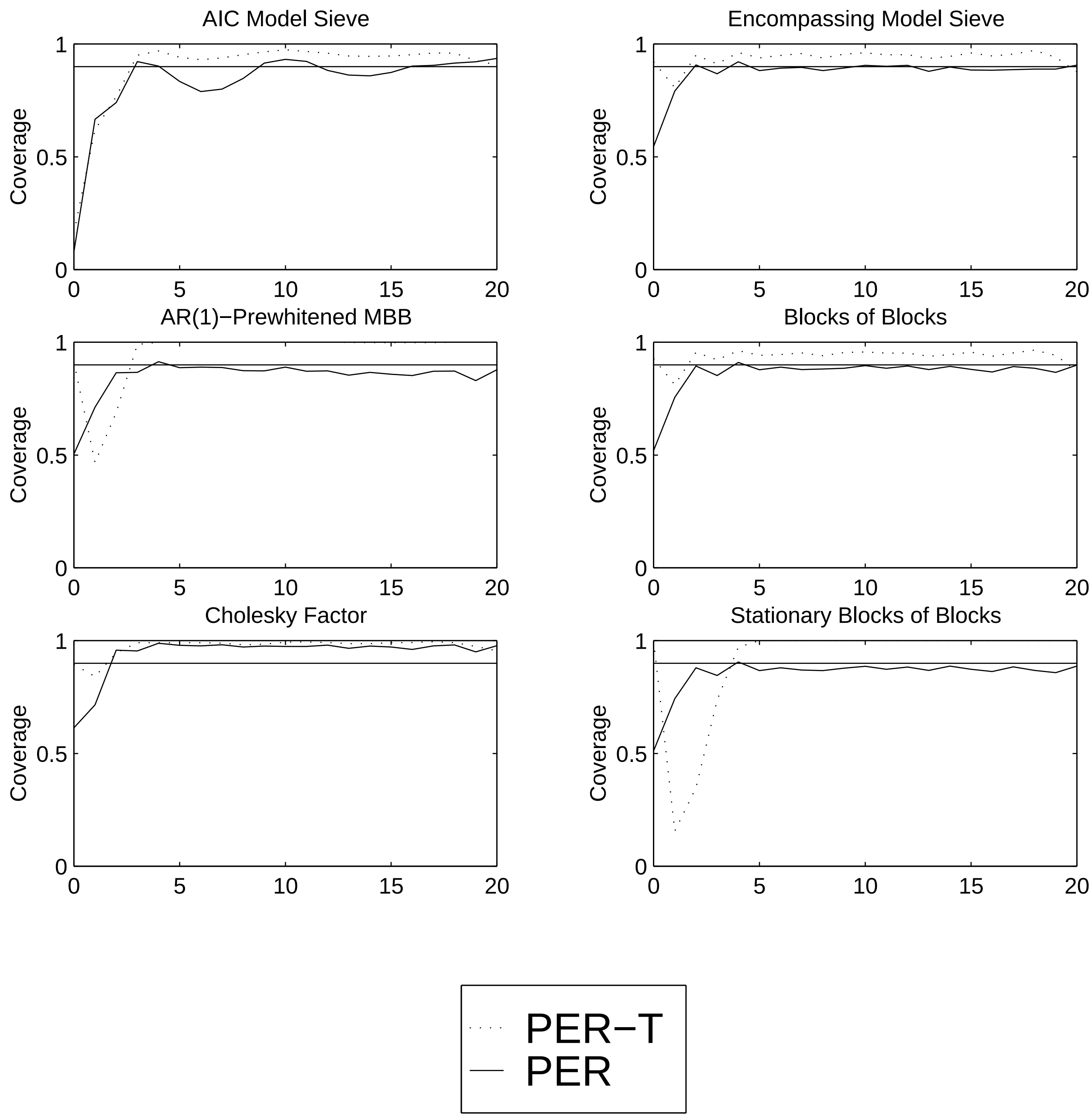


NOTES: See Figure 6.

Figure 8 – Spectral Density

Effective Coverage Rates of Pointwise Nominal 90 % Confidence Intervals for T=240

DGP 3: Growth of Industrial Production

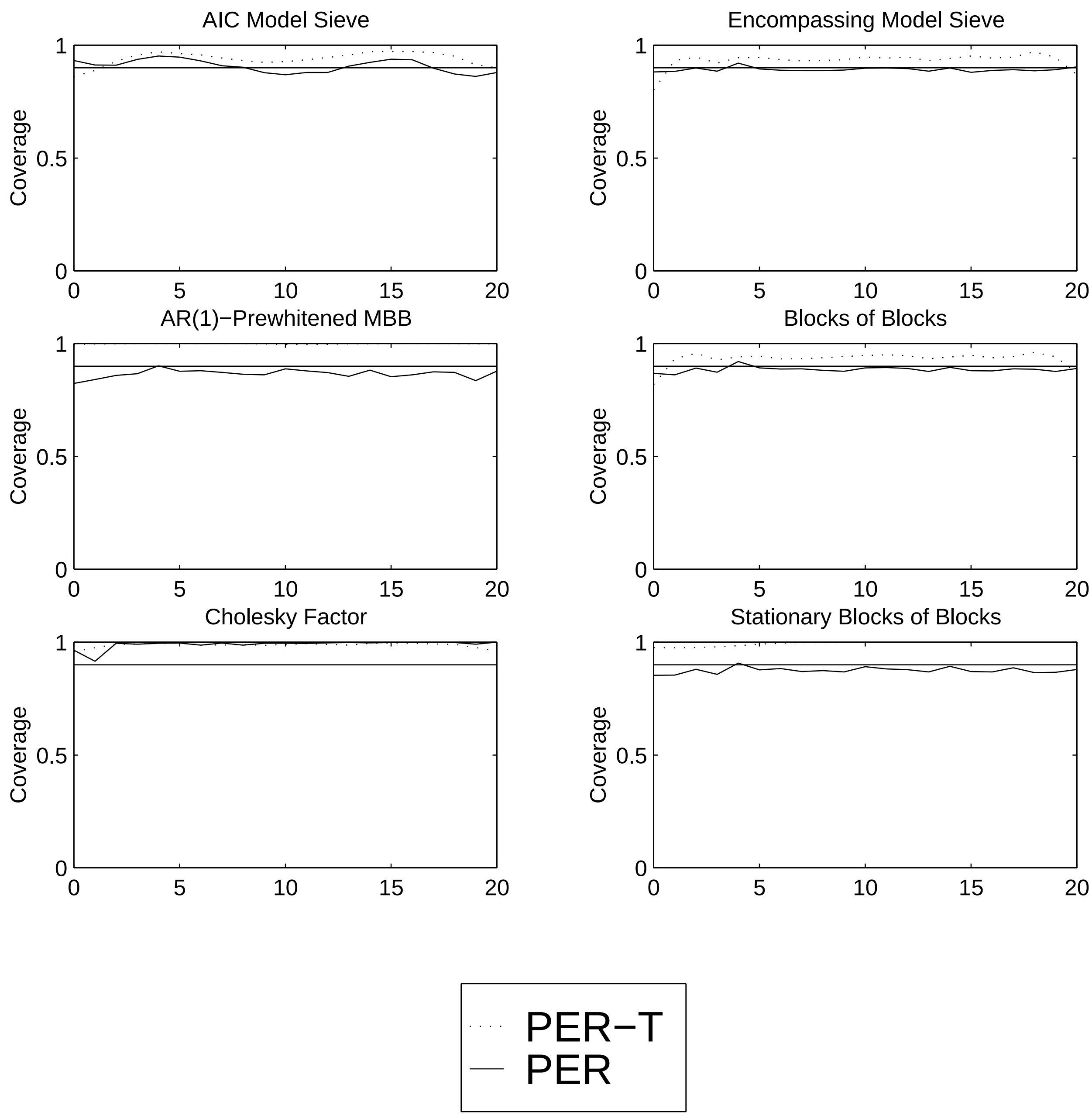


NOTES: See Figure 6.

Figure 9 – Spectral Density

Effective Coverage Rates of Pointwise Nominal 90 % Confidence Intervals for T=240

DGP 4: Percent Change in Yen-Dollar Exchange Rate



NOTES: See Figure 6.