The Finite-Sample Effects of VAR Dimensions on MLE Bias, MLE Variance and Minimum MSE Estimators: Purely Nonstationary Case*

Steve Lawford† and Michalis P. Stamatogiannis
ECARES, Université Libre de Bruxelles, and University of York

March 26, 2002

Abstract

Vector autoregressions (VAR's) are an important tool in time series analysis. However, relatively little is known about the finite-sample behaviour of parameter estimators. We address this issue here, by investigating maximum likelihood estimators (MLE's) in the context of a purely nonstationary first-order VAR. Using Monte Carlo simulation and numerical optimization, we derive response surfaces for MLE bias, in terms of VAR dimensions, given correct and over-parameterization of the model. We study non-zero initial values, and show that univariate bias nonmonotonicity disappears in the multivariate case. Lastly, we examine MLE variance and the correction factors required for the MLE to attain minimum mean squared error (MSE).

Keywords: Finite sample bias; Monte Carlo simulation; Nonstationary time series; Response surfaces; Vector autoregression
JEL classification: C15; C22; C32

*The authors thank Karim Abadir, for his insightful comments on previous versions of this paper; and Walter Distaso, Les Godfrey, Patrick Marsh, and seminar participants at the University of York, for their advice and suggestions. Financial support was provided by a Marie Curie Individual Fellowship under the European Commission's Improving Human Potential programme (Lawford) and a University of York Teaching Fellowship (Stamatogiannis) and is gratefully acknowledged. The authors are solely responsible for information communicated, and the European Community is not responsible for any views or results expressed. This paper was typed in Scientific Word 3.0 and numerical results were derived using GAUSS and E–Views.

†Corresponding author: Steve Lawford, ECARES, Université Libre de Bruxelles, 50 Avenue F. D. Roosevelt, CP 114, B–1050, Brussels, Belgium. (fax: +32 (0)2 650 4475; e-mail: steve_lawford@yahoo.co.uk)

1 Introduction

Vector autoregressions (VAR's) have been extensively studied in econometrics in recent years and continue to be one of the most frequently used tools in time series analysis. They provide a parsimonious means of modelling multiple time series, in which each variable may be related to lags of itself and of all other variables in the system. However, little is currently known about the properties of parameter estimators when applied to finite samples of data, and especially in nonstationary frameworks. In particular, the form and extent of small sample bias have not been fully investigated. This issue is of some practical importance, since small samples and processes with unit roots are typical in many macroeconomic time series; our results will be of interest in this area.[1]

Although the properties of univariate time series were first studied in the 1940's and 1950's by Anderson, Hurwicz, Kendall, Mann, Rubin, Wald and White, among others, and have been thoroughly researched since, important progress on some problems related to the nonstationary case has been made only very recently – derivation of exact results is particularly complicated, even in very simple frameworks, e.g. see Forchini (2001). In the context of nonstationary multiple autoregressive series, Abadir, Hadri and Tzavalis (1999), AHT hereafter, derive an approximate expression for the bias of the maximum likelihood estimator (MLE) of the matrix of autoregressive parameters, in terms of the sample size T and VAR dimension k. They consider estimation of a correctly-parameterized first-order vector autoregression (a VAR(1)), with no constant or trend. Using Monte Carlo simulation, they show that their "analytic approximation" provides a reasonably good representation of bias in finite samples, and for small k (op. cit., p. 166, Table I).

The purposes of our paper are threefold. We extend the results given by AHT for the purely nonstationary (k-dimensional random walk) case in a number of directions, building upon early results by Stamatogiannis (1999) and Lawford (2001, ch. 4). Firstly, we present evidence which suggests that

[1] A huge number of papers have been published in the econometrics and statistics literatures on the subject of vector autoregressive models and modifications thereof, following the influential paper by Sims (1980). For clear introductions to theory and applications, see Lütkepohl (1993, especially chs. 2-3), Johansen (1995, ch. 2) and Ooms (chs. 2-3). For discussion of the use of VAR's in applied macroeconomics, and a comprehensive reference list, see Canova (1995). The unit root framework has also attracted a great deal of attention, e.g. see the important papers by Dickey and Fuller (1979, 1981) and Evans and Savin (1981, 1984). Phillips (1995) provides an interesting discussion of the importance of nonstationarity in economics.

the approximation given by AHT can be improved upon. Subsequently, we use Monte Carlo methods to simulate small sample bias, and then fit a series of response surfaces using nonlinear least squares. A well-specified and parsimonious response surface is chosen following diagnostic checking, and is shown to perform extremely well in out-of-sample forecasting. The forecast error of our response surface is almost always substantially less than that of the AHT form, across the parameter space under investigation; in a number of cases, this error is reduced from more than 10% to below 1%. In broadening the scope of AHT, we also assess over-parameterization of the estimated VAR *model*, (where a constant and where a constant and trend are included). This creates additional bias problems, as was suggested by simulation results for the univariate case in Abadir and Hadri (2000, p. 97) and Tanizaki (2000, p. 505, Table 1). We again propose explanatory functional forms for MLE bias in these frameworks; although these perform less well than that which is derived in the correctly-parameterized case, we note that no other approximations – analytic or otherwise – are currently available in the over-parameterized case.[2]

We motivate the response surface technique as a useful means of improving or "fine-tuning" theoretical results, or of extending them to closely-related frameworks. The difficulty of fitting response surfaces should not be underestimated, even when few parameters are involved. It is important that some prior analytical results should be available, since without such formulae "the search process for the 'best' form would be arbitrary, and would be the equivalent in theoretical econometrics of data mining in applied econometrics" (Abadir, 1993a, p. 88) – for this reason, the response surface methodology is sometimes called a "mixed theory-simulation approach". However, when response surfaces can be found, they are potentially very useful methods of approximation in frameworks where derivation of exact results (although desirable) would be very difficult, or intractable, as is frequently the case even in the simplest settings. Our research may be of use in motivating further theoretical work towards providing rigorous analytical finite-sample proofs. In the meantime, our results can provide guidelines for applied workers as to the behaviour of VAR models given samples sizes that are often relevant in empirical research.

Secondly, we allow the initial values of the VAR data-generating process (DGP) studied by AHT to be non-zero, given a correctly-parameterized

[2]Throughout this paper, we refer to mean-bias as "bias". Median-bias is defined as the difference between the median of an estimator and the true parameter value; see Andrews (1993) for a discussion of median-bias in the context of unit-root/autoregressive models.

model; this setting is likely to be empirically relevant. This follows work by Abadir and Hadri (2000), who investigate the bias of estimated parameters in univariate autoregressive (AR) processes with a (near) unit root, given non-zero initial values, and find that it can be increasing in sample size; a curious effect that is attributable to the non-zero starting values. We show via simulation that this univariate "bias nonmonotonicity" is apparently no longer a problem in the multivariate case, which is a reassuring finding.

Thirdly, we generalize to the VAR framework the univariate results of Abadir (1995a), which we apply directly, once we have noted that our framework is statistically invariant to permutations. We show that multiplying the MLE estimator by a scalar correction factor (which depends here upon T and k) achieves minimum mean squared error (MSE), and removes most of the MLE bias, at the expense of a small increase in estimator variance.

The paper is organized as follows: Section 2 introduces the possibly over-parameterized VAR model, reports the main results of an extensive series of Monte Carlo experiments, and proposes response surfaces for bias. Section 3 deals with the impact of non-zero initial values on bias, in the correctly-parameterized VAR model. Section 4 studies the variance of the MLE for possibly over-parameterized models and examines minimum mean squared error (MSE) estimators. Some concluding remarks are reported in Section 5. Derivations are given in the Appendix.

The notation follows the standard suggested by Abadir and Magnus (2002). We represent scalar, vector and matrix quantities as a, $\boldsymbol{a}$ and $\boldsymbol{A}$ respectively; these have typical elements $\boldsymbol{a} = (\boldsymbol{a})_j$, $\boldsymbol{a}_t = (\boldsymbol{a}_t)_j$ and $\boldsymbol{A} = (\boldsymbol{A})_{ij}$. Special vectors and matrices include the $k \times 1$ zero vector $\boldsymbol{0}_k$, the $k \times k$ zero matrix $\boldsymbol{O}_{kk}$, the $k \times k$ identity matrix $\boldsymbol{I}_k$, and the $k \times 1$ summation vector $\boldsymbol{\imath}_k \equiv (1, 1, \ldots, 1)'$. We denote the expectation operator by $\mathrm{E}(\cdot)$; and $\sum$ signifies a summation from $t = 1$ to T, unless otherwise stated; $\{a_t\}_1^T$ denotes a sequence $a_1, \ldots, a_T$; we use a tilde, e.g. $\widetilde{\psi}$, to represent simulated data. Partial derivatives are written as $\psi_k \equiv \partial\psi/\partial k$. In all Tables except Table 3, $\boldsymbol{x}_0 = \boldsymbol{0}_k$ is assumed.

2 Bias: zero initial values

2.1 Framework

Let $\{\boldsymbol{x}_t\}_1^T$ be a k-dimensional stochastic process that is generated in discrete time by the first-order vector difference equation

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t; \quad t = 1, 2, \ldots T, \tag{1}$$

where the real $k \times k$ matrix of coefficients satisfies

$$\boldsymbol{A} = \boldsymbol{I}_k, \tag{2}$$

and $\{\boldsymbol{\varepsilon}_t\}_1^T \sim \text{IN}(\boldsymbol{0}_k, \boldsymbol{\Omega})$ is a sequence of identically and independently distributed normal random vectors with mean $\boldsymbol{0}_k$ and positive-definite variance-covariance matrix $\boldsymbol{\Omega}$. Rewriting (1) as

$$\boldsymbol{x}_t = \boldsymbol{A}^t \boldsymbol{x}_0 + \boldsymbol{A}^{t-1}\boldsymbol{\varepsilon}_1 + \cdots + \boldsymbol{A}\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{3}$$

we see that the distribution of successive $\boldsymbol{x}_t$'s is not uniquely determined by the $\boldsymbol{\varepsilon}_t$'s; we must also specify the distribution of $\boldsymbol{x}_0$. In this section, we fix $\boldsymbol{x}_0 = \boldsymbol{0}_k$, an initial condition that is often assumed for simplicity in theoretical analysis. We examine the implications of alternative choices for $\boldsymbol{x}_0$ in Section 3. Equations (1) and (2) define a *purely nonstationary* VAR, or vector (Gaussian) random walk, i.e. the variables are unrelated: each $(\boldsymbol{x}_t)_j$ is a function only of its own past value $(\boldsymbol{x}_{t-1})_j$ and an error.

We consider estimation of $\boldsymbol{A}$, given that our maintained model is correctly parameterized, i.e.

$$\boldsymbol{x}_t = \widehat{\boldsymbol{A}}\boldsymbol{x}_{t-1} + \widehat{\boldsymbol{\varepsilon}}_t, \tag{A}$$

which we refer to as Case A. The maximum likelihood estimator $\widehat{\boldsymbol{A}}$ of $\boldsymbol{A}$, based upon an available sample of T observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, is given by

$$\widehat{\boldsymbol{A}} = \left(\sum \boldsymbol{x}_t \boldsymbol{x}_{t-1}'\right)\left(\sum \boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}'\right)^{-1} = \boldsymbol{A} + \left(\sum \boldsymbol{\varepsilon}_t \boldsymbol{x}_{t-1}'\right)\left(\sum \boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}'\right)^{-1}. \tag{4}$$

Given the independent normality of $\boldsymbol{\varepsilon}_t$, (4) coincides with the least-squares estimator of $\boldsymbol{A}$. The bias of $\widehat{\boldsymbol{A}}$ is defined as $\boldsymbol{\Phi} \equiv \text{E}\left(\widehat{\boldsymbol{A}}\right) - \boldsymbol{A}$. In general, the bias will be non-zero in finite samples, and may be quite large – as a result, $\widehat{\boldsymbol{A}}$ can be a particularly misleading indicator of the true $\boldsymbol{A}$ matrix.

2.2 Background

The DGP (1) and (2), with k unit roots, includes as a special case the univariate unit root framework that has been the subject of considerable study. As $k = 1$, the above DGP reduces to the AR(1):

$$x_t = \alpha x_{t-1} + \varepsilon_t; \;\; x_0 = 0; \;\; \alpha = 1; \;\; \varepsilon_t \sim \text{IN}\left(0, \sigma^2\right), \tag{5}$$

and the unrestricted MLE of α is $\widehat{\alpha} = \sum x_t x_{t-1} \left(\sum x_{t-1}^2\right)^{-1}$. The general version of (5) encompasses a wide variety of behaviour, depending upon the choice of parameter values: it may be asymptotically stationary ($|\alpha| < 1$), explosive ($|\alpha| > 1$), or nonstationary ($|\alpha| = 1$; the case $\alpha = 1$ is commonly referred to as a *random walk*). MacKinnon and Smith (1998, p. 207, Figure 1) plot bias functions for $T = 25, 50, 100$, and for $\alpha \in [-1.3, 1.3]$, under univariate Case A ($k = 1$). They show that $\widehat{\alpha}$ is negatively biased for $\alpha = 1$, with bias decreasing in sample size; also, $\widehat{\alpha}$ is positively biased for some values of α (and notably $\alpha = -1$). It can be seen that the bias function for $\widehat{\alpha}$ is almost linear for $\alpha \in [-0.85, 0.85]$, although very nonlinear when $|\alpha| \approx 1$.

The bias and exact moments of the OLS in autoregressive models have been discussed recently by *inter alia* Nankervis and Savin (1988b), Tsui and Ali (1989, 1994), Vinod and Shenton (1996) and Gonzalo and Pitarakis (1998); see also Maeshiro (1999) and Tanizaki (2000) and references therein. Abadir (1993a) derives a high-order closed form (integral-free) analytical approximation for the finite sample bias of $\widehat{\alpha}$ under (5), with $|\alpha| = 1$. The bias of $\widehat{\alpha}$ is given by

$$b \equiv \mathrm{E}\left(\widehat{\alpha}\right) - \alpha = \frac{\sqrt{2}}{T} \mathrm{E}\left(\frac{T\left(\widehat{\alpha} - \alpha\right)}{\sqrt{2}}\right) \equiv \frac{\sqrt{2}}{T} \mathrm{E}\left(z_T\right) \equiv \frac{\sqrt{2}}{T} \mu_T. \tag{6}$$

The distribution of z_T was examined by Evans and Savin (1981) using complex numerical integration. Noting a result of White (1958) – that z_T may be written as a ratio of two quadratic forms in normal variates, with joint moment generating function $M_T(U, V)$ – Abadir (1993a) uses the result $\mu_T = \int_0^\infty \frac{\partial}{\partial U} M_T(U, -V)\,|_{U=0}\, \mathrm{d}V$ to approximate μ_T (and hence b). Dropping terms of order $O\left(T^{-2}\right)$ and above, an expression is derived in terms of nested infinite sums involving parabolic cylinder functions (op. cit., p. 85). This is computable in an efficient manner, using expansions for the confluent hypergeometric functions derived by Abadir (1993b).[3]

The analytical formula used for approximation suggests that exponential functions in polynomials of T^{-1} may be used to describe the bias. A heuristic process gives the simple approximation $\mu_T \approx \mu_\infty \exp\left(-2.6138\, T^{-1}\right)$, where μ_T and μ_∞ are exact values from Evans and Savin (1981, p. 769, Table III). The univariate bias approximation is obtained from (6) and an OLS

[3] Parabolic cylinder functions may also be expressed in terms of Kummer's confluent hypergeometric function, the ${}_1F_1$, although this representation will not necessarily be the most efficient from the point of view of computational generation. See Abadir (1999) for some further discussion of this aspect of special function theory.

regression of $\ln(\mu_T/\mu_\infty)$ on $1/T$ as

$$b \approx -1.7814\left(\frac{1}{T}\right)\exp\left(\frac{-2.6138}{T}\right), \tag{7}$$

where -1.7814 is the expected value of the limiting distribution of $T(\widehat{\alpha}-1)$; e.g. see Le Breton and Pham (1989, p. 562). Despite the fact that only 5 datapoints are used in the derivation of (7), it is accurate to 5 decimal places for bias; and is more accurate than the exact approximation based upon parabolic cylinder functions (see Abadir, 1993a, p. 87, Table 1). Heuristic fits have been used elsewhere in the literature; e.g. Dickey and Fuller (1981, p. 1064). We performed a small study, which compared (7) to simulated bias. This indicated that (7) also works very well out-of-sample, at least to 1 decimal place of $-100\times$bias. It is important to draw a distinction between this "heuristic" fit [5 datapoints, no diagnostics reported] and the response surface fits detailed in Section 2.3 [724 datapoints, diagnostic testing].

In their paper, AHT consider the multivariate version of Case A and provide an approximate formula for $\boldsymbol{\Phi}$, in terms of the sample size and VAR dimension.[4] They prove that $\boldsymbol{\Phi}$ is a scalar matrix, i.e. diagonal with equal diagonal elements: $\boldsymbol{\Phi}=\text{diag}(c,\ldots,c)$; and that $\boldsymbol{\Phi}$ is not a function of $\boldsymbol{\Omega}$. Without any loss of generality, we may then set $\boldsymbol{\Omega}=\boldsymbol{I}_k$ in our simulations. AHT then show that c is approximately equal to k times the univariate bias formula and give the following equation (see AHT, p. 166, and Abadir, 1995a, p. 264):

$$b^{AHT} \approx -1.7814\left(\frac{k}{T}\right)\exp\left(\frac{-2.6138}{T}\right), \tag{AHT1}$$

for $T>k+2$, i.e. bias is proportional to the dimension of the VAR, even when $\boldsymbol{\Omega}=\boldsymbol{I}_k$. It is useful to clarify the circumstances under which this result is valid.

1. If the errors are *symmetric*, and an existence of moments condition is satisfied, then the bias matrix will be exactly a scalar matrix, i.e. $\boldsymbol{\Phi}=c\boldsymbol{I}_k$.

[4] They formulate (1) as $\bigtriangledown\boldsymbol{x}_t=\boldsymbol{C}\boldsymbol{x}_{t-1}+\boldsymbol{\varepsilon}_t$, where the backward-difference operator satisfies $\bigtriangledown\boldsymbol{x}_t\equiv\boldsymbol{x}_t-\boldsymbol{x}_{t-1}$, and $\boldsymbol{C}\equiv\boldsymbol{A}-\boldsymbol{I}_k$. This is primarily to facilitate a subsequent discussion of cointegrating relations. They then study the bias of $\widehat{\boldsymbol{C}}$. Since $\boldsymbol{C}=\boldsymbol{O}_{kk}$ under the DGP, and $\widehat{\boldsymbol{C}}=\widehat{\boldsymbol{A}}-\boldsymbol{I}_k$, we have $\text{E}\left(\widehat{\boldsymbol{C}}\right)-\boldsymbol{C}=\text{E}\left(\widehat{\boldsymbol{A}}\right)-\boldsymbol{A}$. Thus, the bias results reported in this paper may be directly compared to those found in AHT.

2. If the errors are *elliptically symmetric*, then the scalar c is approximately proportional to the univariate bias, i.e. $c \approx k\times$univariate bias.[5] The existence requirement here varies with the density, e.g. $T > k+2$ for Gaussian errors.

3. If the errors are *normal*, the (approximate) univariate bias is given by Abadir (1993a) as (7), and may be substituted to give (AHT1). Explicit formulae for univariate bias are unknown, except for the Gaussian case.

Writing $\psi \equiv \gamma k T^{-1} \exp\left(\beta T^{-1}\right)$, we see that the functional form (AHT1) leads to a number of theoretical predictions with regards to partial derivatives:

$$\psi < 0, \;\; \psi_k < 0, \;\; \psi_T > 0, \;\; \psi_{kk} = 0, \;\; \psi_{kT} > 0, \;\; \psi_{TT} < 0. \tag{8}$$

Given $\gamma = -1.7814$ and $\beta = -2.6138$, these inequalities hold for $T \geq 5$ (see Appendix 6.1). We assess the accuracy of these statements by comparison with simulated data. It is well-known that $\psi < 0$ (e.g. see Figures 1 and 2, which plot simulated bias against sample size and VAR dimension). In fact, most of the statements in (8) agree with the simulated data; see Figures 3 and 4 (details on the simulation experiments are given below). The key statement is $\psi_{kk} = 0$ (see Figure 3, which indicates that $\widetilde{\psi}_{kk} \neq 0$, for T not too large). This provides a motivation for (numerical) refinement of the AHT functional form, e.g. perhaps $\gamma = \gamma(k)$ and $\beta = \beta(k)$ would be more suitable.

2.3 Response surfaces

One of the most useful means of summarizing the results of a (sometimes very large) number of Monte Carlo experiments is estimation of a *response surface*. If the properties of the statistic of interest are simply reported at a number of points in the parameter space, it can become difficult to detect any interesting small-sample behaviour. The response surface technique aims to summarize the behaviour of the statistic of interest at all points in the admissible parameter space, i.e. for whole families of DGP's; and in a more sophisticated manner than that offered by simple heuristic approximations. The following outline of the methodology draws upon Hendry (1984) and Davidson and MacKinnon (1993, pp. 755-763).[6]

[5] See Fang, Kotz and Ng (1990) for a treatment of elliptically symmetric distributions.

[6] Response surfaces are discussed in further detail by Maasoumi and Phillips (1982) and Hendry (1982). The technique is applied in a variety of frameworks by *inter alia* Engle,

The quantity of interest τ is a function of the sample size T and the vector of variables $\boldsymbol{\theta}$ that appear in the DGP. The relationship is modelled as a functional form $\Psi(T, \boldsymbol{\theta}; \boldsymbol{\omega})$, where $\boldsymbol{\omega}$ is a vector of parameters to be estimated, and $\Psi(\cdot)$ is chosen by the investigator. Estimated values for the dependent variable, $\widetilde{\tau}_i$, are generated using a set of N Monte Carlo experiments. The i^{th} experiment is associated with an estimated standard error $\widetilde{\sigma}(\widetilde{\tau}_i)$, where $\widetilde{\tau}_i$ is approximately distributed as $\mathrm{N}\big(\Psi(T, \boldsymbol{\theta}; \boldsymbol{\omega}), \widetilde{\sigma}^2(\widetilde{\tau}_i)\big)$ if the number of replications per experiment (M) is large. Given that each of the experiments uses different sets of random numbers, we may then implement generalized least squares (with a fully specified covariance matrix) and estimate

$$\frac{\widetilde{\tau}_i}{\widetilde{\sigma}(\widetilde{\tau}_i)} = \frac{\Psi(T, \boldsymbol{\theta}; \boldsymbol{\omega})}{\widetilde{\sigma}(\widetilde{\tau}_i)} + \varepsilon_i; \quad \varepsilon_i \sim \mathrm{IN}(0, 1); \quad i = 1, \ldots, N, \tag{9}$$

using ordinary or nonlinear least squares, depending upon the form chosen for $\Psi(T, \boldsymbol{\theta}; \boldsymbol{\omega})$. Division by $\widetilde{\sigma}(\widetilde{\tau}_i)$ in (9) corrects for heteroscedasticity.

There are a number of potential difficulties associated with the approach. Firstly, precise estimates are needed if $\Psi(T, \boldsymbol{\theta}; \boldsymbol{\omega})$ is to be accurately specified. Since a large number (N) of datapoints is also needed – and in practice this seems to be rather more important than having extremely accurate datapoints (although M must be reasonably large) – the method tends to be computationally intensive. Secondly, the functional form of $\Psi(T, \boldsymbol{\theta}; \boldsymbol{\omega})$ is generally not known *a priori*. Thus, estimation of correctly specified response surfaces becomes very difficult indeed as the number of parameters in $\Psi(T, \boldsymbol{\theta}; \boldsymbol{\omega})$ increases. Generally, $\Psi(T, \boldsymbol{\theta}; \boldsymbol{\omega})$ should be formulated in line with known analytical results (as, e.g. we have here in (AHT1)).

Thirdly, Monte Carlo studies can be subject to *specificity* of the results, i.e. while the estimated response surface may fit well in-sample, there is no guarantee that accuracy will be achieved over the entire domain of approximation (Hendry, 1984). To avoid this, T and $\boldsymbol{\theta}$ should be chosen to span an "interesting" part of the parameter space, (e.g. more detail may be given to sample sizes that are typical in economic applications), and the estimated response surface subjected to a battery of standard diagnostic tests. One useful check of the suitability of the response surface specification is that

Hendry and Trumble (1985), Campos (1986), Ericsson (1991) and MacKinnon (1996, 2000). Nankervis and Savin (1988a, especially pp. 129-132) estimate response surfaces for moments of a t statistic in a stationary AR(1), with unknown intercept and normal errors, and compare these with cubic spline interpolation. MacKinnon, Haug and Michelis (1999) and Ericsson and MacKinnon (2001) use response surfaces in their studies of statistics for testing cointegration.

we would expect a unit error variance, after the heteroscedasticity transformation. Inevitably, some (and often a great deal of) experimentation will be required before correctly specified and parsimonious equations can be selected. The accuracy of the approximation should then be examined using out-of-sample parameter values, i.e. points that are not used in estimation of the response surface. This provides a rigorous test of the accuracy of the method and, if the response surface is correctly specified, will enable the statistic of interest to be approximated at various parameter points without the need to carry out another simulation. It is important to report the parameter values used in the simulation experiments; and extreme caution should be exercised when inferring any findings to more general situations than those defined by the DGP and the specific parameter environment (see especially Maasoumi and Phillips, 1982).

In designing our Monte Carlo study, we make use of an indirect simulation technique, known as *antithetic variates*. The method involves finding two different estimates of the quantity of interest in such a way that they will be negatively correlated (or "mutually antithetic"). The information thus obtained is combined to give an improved estimate of the quantity of interest, the degree of increase in accuracy depending upon the strength of the negative correlation. Antithetic variates have been applied in studies of the finite-sample properties of econometric estimators by, e.g. Hendry and Trivedi (1972, especially pp. 119-120) and Mikhail (1972, 1975).

We construct the antithetic variate following Abadir's (1995b) "AV4", which is shown to be useful for the generation of (nearly) nonstationary series (conventional antithetic techniques do not work in our framework). We use the same set of residuals for generation of $\left\{ \boldsymbol{x}_t^- \right\}$, but with parameters of interest in DGP of opposite sign; i.e. $\boldsymbol{x}_t^- = \sum_{j=0}^{t-1} (-\boldsymbol{A})^j \boldsymbol{\varepsilon}_{t-j}$. For illustration, we generated $1,000$ datapoints for Case A, $(T,k) = (25,1)$, using AV4. The antithetic pairs $\left\{ \widehat{\alpha}, \widehat{\alpha}^- \right\}$ have a correlation coefficient of approximately -0.0790 which, although small, gives an effective number of replications M^- of $2,175$; we found this efficiency gain to be quite general. Hence, when generating nonstationary data series, we are able to improve the speed of generation by roughly 50%; or alternatively, we may increase the precision of the resultant estimates, given the same number of replications. This procedure is applied in Case A, for both zero and non-zero initial values; but not for Cases B and C, or for the variance simulations of Section 4.

The bias and variance results reported in the tables were computed using $M = 1,000,000$ replications of the DGP, unless MTk (the total random number requirement) exceeded the period $2^{31} - 1$ of the generator, in which

case $M = 100,000$ was used instead. The simulated values that were used for estimation of response surfaces were generated by choosing the number of replications for a particular T and k such that the Monte Carlo standard error satisfied $\widetilde{\sigma} < 0.0005$. A check of this condition was made every 50 replications, and the simulation stopped when the condition was satisfied. The simulations were carried out on a Pentium 3, 800MHz machine, with 256MB of RAM, using GAUSS 3.2 under Microsoft Windows 2000. Since we chose to adopt a nonlinear $\Psi(\cdot)$, we used the (E–Views) Marquardt optimization algorithm to perform nonlinear least squares. Convergence was very fast, and required few iterations, for a variety of starting values.

The parameters used for estimation of the response surface were:

$$\begin{aligned} T &= 20, 21, \ldots, 200, \\ k &= 1, 2, 3, 4, \end{aligned} \tag{10}$$

which gives $N = 724$ datapoints. The sample sizes that we have chosen are representative of those that are commonly used in practice; and our design includes several small k, so that the effects of changes in VAR dimension can be explored. We calculate the MLE for each combination of (T, k) in the parameter space (10), from which we directly derive the bias. Since $\boldsymbol{\Phi}$ is a scalar matrix, we may estimate the scalar by averaging over the estimated diagonal elements of $\boldsymbol{\Phi}$. This results in a further reduction in the number of replications needed for a given accuracy as k increases.

Following extensive experimentation – based upon (AHT1) and using our observations from the end of Section 2.2 – our estimated response surface is

$$\begin{array}{rcl} b^{RS} & \approx & (0.2617 - 2.0298\,k)\,\dfrac{1}{T}\exp\left(-1.4386\dfrac{1}{T} - 0.8009\,\dfrac{k}{T}\right), \\ & & \begin{array}{cccc} (0.0050) & (0.0020) & (0.0568) & (0.0162) \\ [0.0055] & [0.0022] & [0.0684] & [0.0192] \end{array} \end{array} \tag{11}$$

$\overline{\mathrm{R}}^2 = 0.9999$, S$= 0.9975$, DW$= 1.9037$, JB$= 1.4450$, where $(\cdot)$ are conventional standard errors, $[\cdot]$ are White's (1980) heteroscedasticity-consistent

standard errors, and[7]

$$\begin{aligned}
\overline{\mathrm{R}}^2 &= \text{degrees-of-freedom adjusted coefficient of determination,} \\
\mathrm{S} &= \text{residual standard error,} \\
\mathrm{DW} &= \text{Durbin-Watson } d \text{ test for first-order serial correlation,} \\
\mathrm{JB} &= \text{Jarque and Bera's (1980) test for normality, asymptotically} \\
&\quad \text{distributed as } \chi^2(2).
\end{aligned}$$

We see that the residual standard error is very close to its theoretical value of unity; and the Durbin-Watson d statistic is close to 2, which suggests that there is little or no first-order serial correlation in the estimated errors (the data were first ordered by T and then by k). The Jarque-Bera statistic is not significant, indicating that the estimated errors are well-approximated by the normal distribution.

We recalculate Table I in AHT as Table 1 in this paper, with increased accuracy, with additional results reported for $T = 400, 800$ and $k = 6, 7, 8$, and correcting for a typo in AHT Table I: $(T, k) = (25, 5)$. Clearly, absolute bias is strictly increasing in k and decreasing in T (see also Figures 1 and 2). As T increases, bias goes to zero, as expected from asymptotic theory. We see that b^{AHT} gives a good approximation to bias for k small (and especially for $k = 1$, where (AHT1) reduces to the excellent heuristic approximation derived by Abadir, 1993a). However, as k increases, b^{RS} provides much better results, even for T quite large. Out-of-sample points reported in Table 1 for b^{RS} are combinations of $k = 5, 6, 7, 8$, and $T = 400, 800$. Comparison of the forecast error (f.e.) associated with AHT and our response surface, for various combinations of T and k, indicates that the response surface method leads to considerable improvements; for example, $(T, k) = (25, 8)$: f.e. reduced from 9.9% to 0.1%; $(T, k) = (100, 5)$: f.e. reduced from 7.0% to 0.3%; $(T, k) = (200, 5)$: f.e. reduced from 8.5% to 0.1%; and $(T, k) = (400, 8)$: f.e. reduced from 9.1% to 0.6%. A graphical illustration of the out-of-sample superiority of b^{RS} over b^{AHT} is given in Figure 5, which compares (AHT1) and (11), for $k = 20$ and various T; simulated bias $-100 \times \widetilde{b}$ is plotted as "$\odot$" in the Figure.

We now introduce two further models:

[7] Significance of the statistics at the 5% level of significance is denoted by a superscript $*$ and by $**$ at a 1% level. We follow Urzúa (1996), who notes that the Jarque-Bera test statistic for normality is poorly approximated by its asymptotic $\chi^2(2)$ distribution for small samples. We calculate the finite-sample distribution using Monte Carlo: $M = 100,000$ replications gives 10%, 5% and 1% critical values ($N = 724$) of 4.41, 5.92 and 10.49 respectively. Simulated critical values for DW are given in Appendix 6.2.

$$\boldsymbol{x}_t = \widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{A}}_\mu \boldsymbol{x}_{t-1} + \widehat{\boldsymbol{\varepsilon}}_t \tag{B}$$

$$\boldsymbol{x}_t = \widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\delta}} t + \widehat{\boldsymbol{A}}_{\mu,\delta} \boldsymbol{x}_{t-1} + \widehat{\boldsymbol{\varepsilon}}_t, \tag{C}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$ are $k \times 1$ vectors. These enable us to investigate some of the effects of estimating an incorrectly parameterized model. The MLE's of $\boldsymbol{A}$ in Case B and Case C are denoted by $\widehat{\boldsymbol{A}}_\mu$ and $\widehat{\boldsymbol{A}}_{\mu,\delta}$, where the subscripts μ and μ, δ refer to the fact that a constant and a constant and trend are included in the models. Derivations of $\widehat{\boldsymbol{A}}_\mu$ and $\widehat{\boldsymbol{A}}_{\mu,\delta}$ are given in Appendices 6.3 and 6.4. The bias matrices that correspond to Cases B and C are written as $\boldsymbol{\Phi}_\mu$ and $\boldsymbol{\Phi}_{\mu,\delta}$ respectively.[8]

Theorem 1 *The bias matrices $\boldsymbol{\Phi}_\mu$ and $\boldsymbol{\Phi}_{\mu,\delta}$ are scalar, i.e. $\boldsymbol{\Phi}_\mu \equiv b_\mu \boldsymbol{I}_k$ and $\boldsymbol{\Phi}_{\mu,\delta} \equiv b_{\mu,\delta} \boldsymbol{I}_k$.*

Proof. For a demonstration that $\boldsymbol{\Phi}_{\mu,\delta}$ is scalar, see Appendix 6.5. We do not give details of the proof of this property for $\boldsymbol{\Phi}_\mu$, since it is very similar to (and simpler than) that presented for $\boldsymbol{\Phi}_{\mu,\delta}$. ■

Simulated biases $\widetilde{b}_\mu$ and $\widetilde{b}_{\mu,\delta}$ are reported in Table 2. We see that absolute bias is substantially higher under Case B than under Case A; and higher still for Case C, i.e. incorrect parameterization of the *model* leads to additional bias in MLE estimation. It is very interesting to note that the ratio of $\widetilde{b}_\mu$ to $\widetilde{b}$ is almost constant across T, for given k (e.g. $\widetilde{b}_\mu / \widetilde{b} \approx 3.0$ for $k = 1$ and $\widetilde{b}_\mu / \widetilde{b} \approx 1.3$ for $k = 5$). This relationship is slightly less marked for $\widetilde{b}_{\mu,\delta} / \widetilde{b}$, although a similar comparison can be drawn (e.g. $\widetilde{b}_{\mu,\delta} / \widetilde{b}$ ranges from 2.97 to 3.13 as T ranges from 25 to 800, given $k = 2$). The estimated response surface in Case B is

$$\begin{array}{rcl} b_\mu^{RS} & \approx & \left(-3.5708 - 1.7871\, k - 0.0199\, k^2\right) \dfrac{1}{T} \exp\left(-2.1006 \dfrac{1}{T} - 0.7692\, \dfrac{k}{T}\right), \\ & & \begin{array}{ccccc} (0.0087) & (0.0072) & (0.0015) & (0.0394) & (0.0125) \\ [0.0096] & [0.0076] & [0.0016] & [0.0413] & [0.0153] \end{array} \end{array} \tag{12}$$

$\overline{\mathrm{R}}^2 = 0.9999$, S$= 1.0918$, DW$= 1.7471^{**}$, JB$= 2.5833$. Here, we cannot compare b_μ^{RS} with any analytic approximation, and we see that the response

[8] Of course, $\widehat{\boldsymbol{\mu}}$ is not the same for (B) and (C); neither is $\widehat{\boldsymbol{\varepsilon}}_t$ identical for (A), (B) and (C). However, we do not introduce a special notation, since we are only interested in the maximum likelihood estimators $\widehat{\boldsymbol{A}}$, $\widehat{\boldsymbol{A}}_\mu$ and $\widehat{\boldsymbol{A}}_{\mu,\delta}$.

surface is less well-specified than b^{RS}. For Case C, we estimate the following response surface:

$$b^{RS}_{\mu,\delta} \approx \underset{\substack{(0.0114)\\[0.0152]}}{\left(-8.5818 - 1.6206\,k - 0.0278\,k^2\right)} \frac{1}{T} \exp\left(\underset{\substack{(0.0093)\\[0.0112]}}{-2.9908} \frac{1}{T} \underset{\substack{(0.0019)\\[0.0021]}}{- 0.7304} \frac{k}{T} \right), \quad \underset{\substack{(0.0327)\\[0.0475]}}{} \quad \underset{\substack{(0.0105)\\[0.0200]}}{} \tag{13}$$

$\overline{\mathrm{R}}^2 = 0.9999$, S$= 1.2928$, DW$= 1.3333^{**}$, JB$= 110.2851^{**}$. This response surface is not very well specified. We would not advise the use of either (12) or (13) for approximation out-of-sample.

It is convenient to interpret the scaled bias values reported in Tables 1 and 2 as percentages of the true parameter value; e.g. in Case A, given $(T, k) = (25, 8)$, the absolute bias of each of the estimated parameters on the diagonal of $\widehat{\boldsymbol{A}}$ is 46.7% of the true value (unity). This is despite the fact that the estimated model is correctly parameterized. It is clearly not a good idea to ignore this bias, although its importance from an empirical point of view will depend upon the purpose of the estimate.

One area in which biased estimates may have a detrimental effect is in the estimation of *impulse response functions*, which measure the effect of disturbances ("impulses") at time t on future values of a series, and describe some of the dynamic properties of the system. From (3), it is clear that a temporary (one period) unit shock on $\boldsymbol{\varepsilon}_t$ at time t will have an effect of $\boldsymbol{A}^j$ on $\boldsymbol{x}_t$, j periods in the future, and an estimated impact of $\widehat{\boldsymbol{A}}^j$. Impulse response functions may provide information about the persistence and nature (e.g. direction, asymmetry, peaks) of shocks which, in a general VAR framework, will potentially influence *all* the dependent variables in succeeding periods as lagged values feed through the system.[9]

Under the DGP of pure nonstationarity, a temporary shock will have a permanent effect. The effect of biased estimates can be seen by considering the AR(1) case, e.g. $(T, k) = (25, 1)$, where $\widetilde{b} \approx -0.064$, $\alpha^j = 1$ and $\widehat{\alpha}^j \approx (1 - 0.064)^j$. A unit shock at time t will have an estimated impact at time $t + 5$ of $\widehat{\alpha}^5 \approx 0.7184$. A related concept is the *half-life* h of a shock, defined by $\alpha^h = 0.5$, from which $h = -\ln 2 / \ln \alpha$; this measures how many periods are required for the effect of a temporary unit shock to

[9] An interesting question that arises in economics is whether demand and supply disturbances will have a permanent or temporary impact upon some dependent variable. Related studies that incorporate estimation of impulse response functions include Backus (1986, pp. 634-636) and Blanchard and Quah (1989, pp. 662-664).

be reduced by 50%. As $\alpha \to 1$, $h \to \infty$ (corresponding to a permanent effect). However, for $\widehat{\alpha} \approx 1 - 0.064 = 0.936$, $h \approx 10.48$ periods. Bias leads to an underestimation of both the estimated impact and the half-life of a temporary shock (see Patterson, 2000, for an illustration of this problem in a stationary autoregressive setting).

3 Bias: non-zero initial values

> "We generally believe that the more information we use the better the estimator will be and this is certainly a characteristic we would wish estimators to have. The monotonicity properties of estimators do not seem to have attracted much attention in either the statistical or econometric literature. Perhaps we should be mindful of such considerations when making a choice amongst competing estimators." (Ip and Phillips, 1998, p. 309)

We investigate by simulation how the results of Section 2 differ when we remove the restriction on initial values, and allow $\boldsymbol{x}_0 \neq \mathbf{0}_k$. Specifically, we assess Case A given correct parameterization of the model, since it is in the DGP with no deterministic components that the distribution of $\widehat{\mathbf{A}}$ depends upon the initial condition. This setting was examined by Abadir and Hadri (2000), for the driftless AR(1), with a root close to (and below) unity. They perform the same numerical integration as Evans and Savin (1981, especially p. 769), to calculate their Tables 1-4 (op. cit., p. 95), which report exact bias, up to an integration error. They make the disquieting observation that the bias of the estimated autoregressive parameter can *increase* with sample size T; an effect known as *bias nonmonotonicity*. They also show that bias is not a monotonic function of the autoregressive root.

Bias nonmonotonicity here is due to the effect of the initial observations $\boldsymbol{x}_0$: in the univariate case, this arises as $\sigma^{-1}|x_0|$ increases. This phenomenon was first reported by Abadir (1993a, p. 90) for $|\alpha| = 1$; and then by Abadir and Hadri (2000), for $\alpha = 0.90, 0.95, 0.99, 1.00$; as $\alpha \to 0$, the initial value x_0 becomes irrelevant, as is well-known. In Table 3, we report $-100 \times \widetilde{b}$ for $T = 25, 50, 100, 200$ and $k = 2, 3, 4, 5, 6$; and $\boldsymbol{x}_0'\boldsymbol{x}_0 = 10, 20, 50, 100, 150$. For a given value $\boldsymbol{x}_0'\boldsymbol{x}_0$, various choices of $\boldsymbol{x}_0$ are possible: we generated elements of $\boldsymbol{x}_0$ as $(\boldsymbol{x}_0)_j = +\sqrt{\boldsymbol{x}_0'\boldsymbol{x}_0/k}$, for $j = 1, 2, \ldots, k$. It is important to note that we chose $\boldsymbol{x}_0'\boldsymbol{x}_0$ arbitrarily; it is a simple scalar function of $\boldsymbol{x}_0$ that enables us to *illustrate* the effect of non-zero initial values.

Abadir and Hadri (2000) indicate that overparameterization of the model in the univariate setting reverses bias nonmonotonicity, at the expense of

higher bias everywhere. From Table 3 here, it also seems clear that bias nonmonotonicity disappears as we move to a multivariate framework (e.g. $\boldsymbol{x}_0'\boldsymbol{x}_0 = 150$; $k = 2$; $-100 \times \widetilde{b} = 9.8, 5.3, 2.9, 1.5$ for $T = 25, 50, 100, 200$). A formal proof of this conjecture is as yet unavailable in the literature.[10]

4 Variance and MSE

The $m \times m$ variance-covariance matrix of an $m \times 1$ vector $\boldsymbol{r}$ is defined as $\boldsymbol{V}(\boldsymbol{r}) \equiv \mathrm{E}\left[(\boldsymbol{r} - \mathrm{E}(\boldsymbol{r}))(\boldsymbol{r} - \mathrm{E}(\boldsymbol{r}))'\right]$. If $\widehat{\boldsymbol{A}}$ has typical element $\widehat{a}_{rs}$, then the scalar variance of $\widehat{a}_{rs}$ will be located as element v_{qq} of the $k^2 \times k^2$ variance-covariance matrix of $\mathrm{vec}\left(\widehat{\boldsymbol{A}}\right)$, where $q = (s-1)k + r$, for $r, s = 1, 2, \ldots, k$. When $r = s$ (corresponding to diagonal elements of $\widehat{\boldsymbol{A}}$) we have $q = (r-1)(k+1) + 1$.

Since our framework is statistically invariant to permutations, the variance of each of the diagonal elements of $\widehat{\boldsymbol{A}}$ is identical; we denote the scalar variance as V. Simulated values of $\widetilde{V}$ are given in Table 4, for various T, k and for Cases A, B and C. Clearly, variance is increasing in k and decreasing in T; and also increases as the model is overparameterized.[11] This latter result, combined with higher bias under Cases B and C, shows the importance of parsimonious modelling in econometrics.

Abadir (1995a) use the definition

$$V \equiv \mathrm{E}\left(\left\{\widehat{\alpha} - \mathrm{E}(\widehat{\alpha})\right\}^2\right) = \frac{2}{T^2}\mathrm{E}\left(\left\{z_T - \mathrm{E}(z_T)\right\}^2\right) \equiv \frac{2}{T^2}\mathrm{var}(z_T)$$

in the univariate framework, with values for standard deviation "SD" for normalized $\widehat{\alpha}$ taken from Evans and Savin (1981, p. 769, Table III), where $\mathrm{V}=\frac{2}{T^2}\mathrm{SD}^2$; and performs a similar heuristic process to that used in derivation of (7) for bias. This gives (op. cit., p. 265)

$$V \approx \frac{10.1124}{T^2}\exp\left(\frac{-5.4462}{T} + \frac{14.519}{T^2}\right), \tag{14}$$

[10] Bias nonmonotonicity is discovered in a different setting by Ip and Phillips (1998), who show that neither the exact bias nor the exact MSE of the two-stage least squares exogenous variable coefficient estimators in the structural equations of a two equation static simultaneous equation model are monotonically non-increasing with sample size. Nonmonotonicity is shown for *small* samples, e.g. $\mathrm{n} \leq 30, 50$, for unit increases in sample size. Their results are not the result of the choice of starting value; neither do we find nonmonotonicity of MSE in our framework.

[11] The simulated variance of the *off-diagonal* elements of $\widehat{\boldsymbol{A}}$ is of a similar magnitude to that of V.

which is shown to be accurate to at least 7 decimal places in small samples. Bias and variance are not the only criteria to be used in comparison of time series estimates. Using (7) and (14), and $\text{MSE}(\widehat{\alpha}) = b^2 + V$, Abadir (1995a) derives the mean squared error corresponding to our correctly parameterized univariate Case A. Defining $\lambda = \lambda(T)$ as a correction factor such that $\text{MSE}(\lambda\widehat{\alpha})$ is minimized, it is shown that (op. cit., p. 265)

$$\lambda = \frac{1+b}{V+(1+b)^2} \tag{15}$$

when $\alpha = 1$; we see that $\lambda \to 1$ as T increases. It is then possible to substitute in exact or simulated values for bias and variance in order to calculate λ for various T (see op. cit., p. 266, Table 2). Figure 6 plots $\text{MSE}(\lambda\widehat{\alpha})$ against the correction factor λ, for $k = 1$ and $T = 25, 50, 100$. It is clear that MLE ($\lambda = 1$) does not achieve minimum MSE. Also, for $\alpha = 1$,

$$b^m = \frac{-V}{V+(1+b)^2}; \quad V^m = \lambda^2 V \geq V, \tag{16}$$

where b^m and V^m denote bias and variance of the *corrected* MLE; it is shown that the corrected (for minimum MSE) MLE is almost unbiased, unlike the MLE.

In our multivariate setting, we are able to substitute simulated $\widetilde{b}$ and $\widetilde{V}$ into (15) and (16) directly for $k > 1$, in order to calculate Tables 5, 6 and 7. Although the only difference is that $\lambda = \lambda(T, k)$ is now a function of the VAR dimension, these results have never been reported in the literature. We see from Table 5 that the correction factor increases with k and falls with T; asymptotically, the MLE achieves minimum MSE. The correction λ can be particularly large for small T, e.g. $(T, k) = (50, 5)$ implies a correction of 19%. Tables 6 and 7 display qualitatively similar results to those in Abadir (1995a), i.e. the corrected estimator $\lambda\widehat{\boldsymbol{A}}$ is much less biased than the MLE, and b^m tends to zero much more rapidly than b; however, this reduction in bias comes at the expense of a small increase in the variance of the corrected estimator, V^m.

Tanizaki (2000) develops a simulation-based bias correction procedure and reports (op. cit., p. 505, Table 1) results that correspond to a mean-bias correction in our univariate Case A, for $T = 19, 39, 59$: he gives simulated $-100\times$bias as approximately $7.8, 4.0$ and 2.8 for these sample sizes, with mean-bias corrected $-100\times$bias of $0.7, 0.2$ and 0.3 [this apparent non-monotonicity arises since only $M = 10,000$ replications are used]. Using our method for minimizing MSE, and $M = 1,000,000$ replications, we calculated $-100 \times \widetilde{b}$ and $-100 \times b^m$ for $k = 1$ and $T = 19, 39, 59$ as $8.2, 4.3, 2.9$;

and $2.5, 0.6, 0.3$ respectively. Our method for minimizing MSE therefore has the added benefit of a substantial bias reduction, even when compared with an actual bias-reduction technique.

5 Concluding comments

We have reported a wide variety of simulation results for the bias and variance of the MLE of the autoregressive parameters, given a purely nonstationary VAR, for small T and k; and where the estimated model is possibly overparameterized. Although special cases of our general framework have been the subject of considerable research in econometrics, a comprehensive simulation study of this nature has not previously been performed. Wherever possible, our numerical results were cross-checked with partial exact and approximate results in the literature, e.g. Pere (2000, p. 352, Table 3) reports values that correspond to variances in our univariate Case B, $T = 25, 100$, in his study of adjusted profile likelihood.

We have extended recent work in papers by Abadir, Hadri and Tzavalis. We successfully estimate a simple, parsimonious and well-specified response surface that refines an analytical result of Abadir, Hadri and Tzavalis (1999), and considerably improves upon its predictive ability. Although it would be desirable to find exact expressions for density functions and moment formulae in our frameworks, these are currently intractable and will probably involve multiple infinite series of matrix-argument hypergeometric functions (generalizing, e.g. Abadir, 1993a). Exact formulae give rise to series related to hypergeometric functions in other areas, e.g. see Phillips (1980) and references therein – they are generally very complicated and may be difficult to implement for numerical evaluation. We may, therefore, need to rely upon approximations even when the exact formulae are available.

We examine bias and variance, given overparameterization of the estimated model. This provides a strong motivation for parsimonious modelling in econometrics. We also relax the condition on initial values, $\boldsymbol{x}_0 = \mathbf{0}_k$, and use our simulation results to propose the following conjecture: univariate bias nonmonotonicity disappears as we move to the multivariate setting. We tabulate the correction factors required for the MLE to achieve minimum MSE and show that this correction can significantly reduce bias, at the expense of a small increase in estimator variance.

Our work complements asymptotic treatments by *inter alia* Phillips (1987a, 1987b) in the univariate framework; and Phillips and Durlauf (1986), Park and Phillips (1988, 1989) and Tsay and Tiao (1990) in the multivari-

ate setting. It is our hope that the results presented in this paper will help applied econometricians to understand the properties of the MLE given sample sizes that are useful in practical work. Theoretical econometricians may find them helpful when studying the derivation of exact formulae, e.g. in conjunction with work by Abadir and Larsson (1996, 2001), who derive the exact finite sample moment generating function of the quadratic forms that create the basis for the sufficient statistic in a discrete Gaussian vector autoregression. We are currently investigating this latter issue.

6 Appendix

6.1 Properties of AHT Theorem 1

From (AHT1), we define the general form $\psi \equiv \gamma k T^{-1} \exp\left(\beta T^{-1}\right)$, with $\gamma, \beta < 0$. Taking partial derivatives, we have the following:

$$\begin{aligned}
\psi_k &= \gamma T^{-1} \exp\left(\beta T^{-1}\right) \\
\psi_T &= -\gamma k T^{-3} (\beta + T) \exp\left(\beta T^{-1}\right) \\
\psi_{kk} &= 0 \\
\psi_{kT} &= -\gamma T^{-3} (\beta + T) \exp\left(\beta T^{-1}\right) \\
\psi_{TT} &= \gamma k T^{-5} \left(\beta^2 + 4\beta T + 2T^2\right) \exp\left(\beta T^{-1}\right).
\end{aligned}$$

We note that ψ is infinitely differentiable with respect to T. Now, it is clear that $\psi < 0$, $\psi_k < 0$ and $\psi_{kk} = 0$. Also, $\psi_T > 0$ and $\psi_{kT} > 0$ as $T > -\beta$. Finally, $\psi_{TT} > 0$ as $\beta^2 + 4\beta T + 2T^2 < 0$, i.e. as $\left(-2 - \sqrt{2}\right) T < \beta < \left(-2 + \sqrt{2}\right) T$. For $\beta = -2.6138$, it follows that $\psi_T > 0$ and $\psi_{kT} > 0$ as $T \geq 3$. Also, $\beta > \left(-2 + \sqrt{2}\right) T$, and therefore $\psi_{TT} < 0$, as $T \geq 5$.

6.2 Durbin-Watson test for serial correlation

Given that there is no intercept in the regression, then Durbin and Watson's d statistic is bounded as $d \in [d_M, d_U]$, where the lower and upper bounding distributions are defined by (Farebrother, 1980)

$$d_M = \frac{\sum_{i=1}^{n-k} \lambda_i \zeta_i^2}{\sum_{i=1}^{n-k} \zeta_i^2}; \quad d_U = \frac{\sum_{i=1}^{n-k} \lambda_{i+k} \zeta_i^2}{\sum_{i=1}^{n-k} \zeta_i^2},$$

where $\lambda_i = 2 - 2\cos\{\pi (i-1)/n\} \in [0, 4)$ and the ζ_i's are independent $\mathrm{N}(0,1)$ variates; $i = 1, 2, \ldots, n$. Using $100,000$ replications, we simulated

the following tail percentiles of d_M and d_U, for $n = 724$ and $k = 4$ ($k = 5$ in parentheses):

	1%	5%	95%	99%
d_M	1.81 (1.81)	1.86 (1.86)	2.11 (2.11)	2.16 (2.16)
d_U	1.84 (1.84)	1.89 (1.89)	2.13 (2.14)	2.18 (2.19)

6.3 Form of $\widehat{\boldsymbol{A}}_\mu$ in Case B

We begin with

$$\boldsymbol{x}_t = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t \equiv \boldsymbol{\Psi}\boldsymbol{z}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\Psi} \equiv (\boldsymbol{\mu}, \boldsymbol{A})$ and $\boldsymbol{z}'_{t-1} \equiv (1, \boldsymbol{x}'_{t-1})$ are defined. Then, $\widehat{\boldsymbol{\Psi}} = \boldsymbol{\Psi} + \left(\sum \boldsymbol{\varepsilon}_t \boldsymbol{z}'_{t-1}\right)\left(\sum \boldsymbol{z}_{t-1}\boldsymbol{z}'_{t-1}\right)^{-1}$ and

$$\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{A}}_\mu\right) = (\boldsymbol{\mu}, \boldsymbol{A}) + \left(\sum \boldsymbol{\varepsilon}_t, \sum \boldsymbol{\varepsilon}_t \boldsymbol{x}'_{t-1}\right) \begin{pmatrix} T & \sum \boldsymbol{x}'_{t-1} \\ \sum \boldsymbol{x}_{t-1} & \sum \boldsymbol{x}_{t-1}\boldsymbol{x}'_{t-1} \end{pmatrix}^{-1}.$$

If the block-diagonal elements of $\sum \boldsymbol{z}_{t-1}\boldsymbol{z}'_{t-1}$ are invertible, and the following terms exist:

$$\begin{aligned} d_1 &= \left[T - \left(\sum \boldsymbol{x}'_{t-1}\right)\left(\sum \boldsymbol{x}_{t-1}\boldsymbol{x}'_{t-1}\right)^{-1}\left(\sum \boldsymbol{x}_{t-1}\right)\right]^{-1} \\ \boldsymbol{D}_3 &= \left[\sum \boldsymbol{x}_{t-1}\boldsymbol{x}'_{t-1} - \frac{1}{T}\left(\sum \boldsymbol{x}_{t-1}\right)\left(\sum \boldsymbol{x}'_{t-1}\right)\right]^{-1} \\ \boldsymbol{d}_2 &= -\frac{1}{T}\boldsymbol{D}_3\left(\sum \boldsymbol{x}_{t-1}\right), \end{aligned}$$

we may implement the formula for partitioned inverses; e.g. Magnus and Neudecker (1999, pp. 11-12). Clearly, T^{-1} exists for all $T \neq 0$. The symmetric matrix $\sum \boldsymbol{x}_{t-1}\boldsymbol{x}'_{t-1}$ becomes just invertible, almost surely (with probability 1) when there are as many nonzero terms in the sum as dimensions, these being generated in a non-collinear fashion (errors are i.i.d.). For $\boldsymbol{x}_0 = \boldsymbol{0}_k$, this holds as $T > k$. It then follows that

$$\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{A}}_\mu\right) = (\boldsymbol{\mu}, \boldsymbol{A}) + \left(\sum \boldsymbol{\varepsilon}_t, \sum \boldsymbol{\varepsilon}_t \boldsymbol{x}'_{t-1}\right) \begin{pmatrix} d_1 & \boldsymbol{d}'_2 \\ \boldsymbol{d}_2 & \boldsymbol{D}_3 \end{pmatrix}.$$

Since $\left(\sum \boldsymbol{x}_{t-1}\right)' = \left(\sum \boldsymbol{x}'_{t-1}\right)$ and $\boldsymbol{D}_3$ is symmetric, we have

$$\begin{aligned}\widehat{\boldsymbol{A}}_{\mu} &= \boldsymbol{A} + \left(\sum \boldsymbol{\varepsilon}_t\right)\boldsymbol{d}_2' + \left(\sum \boldsymbol{\varepsilon}_t \boldsymbol{x}'_{t-1}\right)\boldsymbol{D}_3 \\ &= \boldsymbol{A} + \left[\sum \boldsymbol{\varepsilon}_t \boldsymbol{x}'_{t-1} - \frac{1}{T}\left(\sum \boldsymbol{\varepsilon}_t\right)\left(\sum \boldsymbol{x}'_{t-1}\right)\right]\boldsymbol{D}_3.\end{aligned}$$

6.4 Form of $\widehat{\boldsymbol{A}}_{\mu,\delta}$ in Case C

We have

$$\boldsymbol{x}_t = \boldsymbol{\mu} + \boldsymbol{\delta} t + \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t \equiv \boldsymbol{\Theta}\boldsymbol{w}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\Theta} \equiv (\boldsymbol{E}, \boldsymbol{A})$, $\boldsymbol{E} \equiv (\boldsymbol{\mu}, \boldsymbol{\delta})$, $\boldsymbol{w}'_{t-1} \equiv (\boldsymbol{q}', \boldsymbol{x}'_{t-1})$ and $\boldsymbol{q}' \equiv (1, t)$. Then, $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta} + \left(\sum \boldsymbol{\varepsilon}_t \boldsymbol{w}_{t-1}\right)\left(\sum \boldsymbol{w}_{t-1}\boldsymbol{w}'_{t-1}\right)^{-1}$. Hence,

$$\left(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{A}}_{\mu,\delta}\right) = (\boldsymbol{E}, \boldsymbol{A}) + \left(\sum \boldsymbol{\varepsilon}_t \boldsymbol{q}', \sum \boldsymbol{\varepsilon}_t \boldsymbol{x}'_{t-1}\right)\begin{pmatrix} \sum \boldsymbol{q}\boldsymbol{q}' & \sum \boldsymbol{q}\boldsymbol{x}'_{t-1} \\ \sum \boldsymbol{x}_{t-1}\boldsymbol{q}' & \sum \boldsymbol{x}_{t-1}\boldsymbol{x}'_{t-1} \end{pmatrix}^{-1}.$$

Note that

$$\sum \boldsymbol{q}\boldsymbol{q}' = \begin{pmatrix} T & \sum t \\ \sum t & \sum t^2 \end{pmatrix} = \begin{pmatrix} T & \frac{1}{2}T(T+1) \\ \frac{1}{2}T(T+1) & \frac{1}{6}T(T+1)(2T+1) \end{pmatrix},$$

from which the determinant $\det\left(\sum \boldsymbol{q}\boldsymbol{q}'\right) = \frac{1}{12}T^2(T+1)(T-1)$. Thus, $\det\left(\sum \boldsymbol{q}\boldsymbol{q}'\right) \geq 0$ as $T \geq 1$, with equality at $T = 1$; and $\left(\sum \boldsymbol{q}\boldsymbol{q}'\right)^{-1}$ exists for $T > 1$. Given existence of $\left(\sum \boldsymbol{x}_{t-1}\boldsymbol{x}'_{t-1}\right)^{-1}$ and the following:

$$\begin{aligned}\boldsymbol{F}_1 &= \left[\sum \boldsymbol{q}\boldsymbol{q}' - \left(\sum \boldsymbol{q}\boldsymbol{x}'_{t-1}\right)\left(\sum \boldsymbol{x}_{t-1}\boldsymbol{x}'_{t-1}\right)^{-1}\left(\sum \boldsymbol{x}_{t-1}\boldsymbol{q}'\right)\right]^{-1} \\ \boldsymbol{F}_3 &= \left[\sum \boldsymbol{x}_{t-1}\boldsymbol{x}'_{t-1} - \left(\sum \boldsymbol{x}_{t-1}\boldsymbol{q}'\right)\left(\sum \boldsymbol{q}\boldsymbol{q}'\right)^{-1}\left(\sum \boldsymbol{q}\boldsymbol{x}'_{t-1}\right)\right]^{-1} \\ \boldsymbol{F}_2 &= -\boldsymbol{F}_3\left(\sum \boldsymbol{x}_{t-1}\boldsymbol{q}'\right)\left(\sum \boldsymbol{q}\boldsymbol{q}'\right)^{-1},\end{aligned}$$

then

$$\left(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{A}}_{\mu,\delta}\right) = (\boldsymbol{E}, \boldsymbol{A}) + \left(\sum \boldsymbol{\varepsilon}_t \boldsymbol{q}', \sum \boldsymbol{\varepsilon}_t \boldsymbol{x}'_{t-1}\right)\begin{pmatrix} \boldsymbol{F}_1 & \boldsymbol{F}_2' \\ \boldsymbol{F}_2 & \boldsymbol{F}_3 \end{pmatrix}.$$

Directly, for $T > k$, and noting that $\left(\sum \boldsymbol{q}\boldsymbol{q}'\right)^{-1}$ and $\boldsymbol{F}_3$ are symmetric, and that $\left(\sum \boldsymbol{x}_{t-1}\boldsymbol{q}'\right)'=\sum \boldsymbol{q}\boldsymbol{x}_{t-1}'$,

$$\widehat{\boldsymbol{A}}_{\mu,\delta} = \boldsymbol{A} + \left[\sum \boldsymbol{\varepsilon}_t \boldsymbol{x}_{t-1}' - \left(\sum \boldsymbol{\varepsilon}_t \boldsymbol{q}'\right)\left(\sum \boldsymbol{q}\boldsymbol{q}'\right)^{-1}\left(\sum \boldsymbol{q}\boldsymbol{x}_{t-1}'\right)\right]\boldsymbol{F}_3. \quad (17)$$

6.5 Proof of Theorem 1 for Case C

The proof of this proposition uses the same symmetry/antithetic arguments as AHT (p. 165), for Case A, where $\boldsymbol{C} \equiv A - \boldsymbol{I}_k$. The corresponding proof for Case B is very similar to this one and we do not include it here. Define $\boldsymbol{e}_t \equiv \boldsymbol{L}^{-1}\boldsymbol{u}_t$, $\boldsymbol{z}_t \equiv \boldsymbol{L}^{-1}\boldsymbol{x}_t$, where $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{L}'$ from the Cholesky decomposition and $\boldsymbol{L}$ is a $k \times k$ lower-triangular matrix. Then, from (17), we have

$$\begin{aligned}\widehat{\boldsymbol{C}}_{\mu,\delta} &= \boldsymbol{C} + \boldsymbol{L}\left[\sum \boldsymbol{e}_t \boldsymbol{z}_{t-1}' - \left(\sum \boldsymbol{e}_t \boldsymbol{q}'\right)\left(\sum \boldsymbol{q}\boldsymbol{q}'\right)^{-1}\left(\sum \boldsymbol{q}\boldsymbol{z}_{t-1}'\right)\right] \times \\ &\quad \left[\sum \boldsymbol{z}_{t-1}\boldsymbol{z}_{t-1}' - \left(\sum \boldsymbol{z}_{t-1}\boldsymbol{q}'\right)\left(\sum \boldsymbol{q}\boldsymbol{q}'\right)\left(\sum \boldsymbol{q}\boldsymbol{z}_{t-1}'\right)\right]\boldsymbol{L}^{-1},\end{aligned}$$

where we (implicitly) define the above product so that $\widehat{\boldsymbol{C}}_{\mu,\delta} = \boldsymbol{C} + \boldsymbol{L}\boldsymbol{R}\boldsymbol{L}^{-1}$, from which $\boldsymbol{\Phi}_{\mu,\delta} \equiv \mathrm{E}\left(\widehat{\boldsymbol{C}}_{\mu,\delta}\right) = \boldsymbol{L}\,\mathrm{E}\left(\boldsymbol{R}\right)\boldsymbol{L}^{-1}$. We now define

$$\begin{aligned}\boldsymbol{P} &\equiv \sum \boldsymbol{e}_t \boldsymbol{z}_{t-1}' = \begin{bmatrix} p_1 & \boldsymbol{p}_2' \\ \boldsymbol{p}_3 & \boldsymbol{P}_4 \end{bmatrix} \\ \boldsymbol{Q} &\equiv \sum \boldsymbol{z}_{t-1}\boldsymbol{z}_{t-1}' = \begin{bmatrix} q_1 & \boldsymbol{q}_2' \\ \boldsymbol{q}_3 & \boldsymbol{Q}_4 \end{bmatrix} \\ \boldsymbol{W} &\equiv \sum \boldsymbol{q}\boldsymbol{q}' = \begin{bmatrix} w_1 & w_2 \\ w_3 & w_4 \end{bmatrix},\end{aligned}$$

and denote inverse elements using superscripts. We can then write

$$\begin{aligned}\boldsymbol{R} \equiv &\left(\begin{bmatrix} p_1 & \boldsymbol{p}_2' \\ \boldsymbol{p}_3 & \boldsymbol{P}_4 \end{bmatrix} - \left[\sum \boldsymbol{e}_t \vdots \sum t\boldsymbol{e}_t\right]\begin{bmatrix} w^{(1)} & w^{(2)} \\ w^{(3)} & w^{(4)} \end{bmatrix}\begin{bmatrix} \sum \boldsymbol{z}_{t-1}' \\ \sum t\boldsymbol{z}_{t-1}' \end{bmatrix}\right) \times \\ &\left(\begin{bmatrix} q_1 & \boldsymbol{q}_2' \\ \boldsymbol{q}_3 & \boldsymbol{Q}_4 \end{bmatrix} - \left[\sum \boldsymbol{z}_{t-1} \vdots \sum t\boldsymbol{z}_{t-1}\right]\begin{bmatrix} w_1 & w_2 \\ w_3 & w_4 \end{bmatrix}\begin{bmatrix} \sum \boldsymbol{z}_{t-1}' \\ \sum t\boldsymbol{z}_{t-1}' \end{bmatrix}\right). \quad (18)\end{aligned}$$

We then define

$$\begin{aligned}\boldsymbol{D} \equiv &\begin{bmatrix} d_1 & \boldsymbol{d}_2' \\ \boldsymbol{d}_3 & \boldsymbol{D}_4 \end{bmatrix} = w^{(1)}\left(\sum \boldsymbol{e}_t\right)\left(\sum \boldsymbol{z}_{t-1}'\right) + w^{(2)}\left(\sum \boldsymbol{e}_t\right)\left(\sum t\boldsymbol{z}_{t-1}'\right) \\ + &\; w^{(3)}\left(\sum t\boldsymbol{e}_t\right)\left(\sum \boldsymbol{z}_{t-1}'\right) + w^{(4)}\left(\sum t\boldsymbol{e}_t\right)\left(\sum t\boldsymbol{z}_{t-1}'\right)\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{E} &\equiv \left[\begin{array}{cc} e_1 & \boldsymbol{e}_2' \\ \boldsymbol{e}_3 & \boldsymbol{E}_4 \end{array}\right] = w_1 \left(\sum \boldsymbol{z}_{t-1}\right)\left(\sum \boldsymbol{z}_{t-1}'\right) + w_2 \left(\sum \boldsymbol{z}_{t-1}\right)\left(\sum t\boldsymbol{z}_{t-1}'\right) \\ &+ w_3 \left(\sum t\boldsymbol{z}_{t-1}\right)\left(\sum \boldsymbol{z}_{t-1}'\right) + w_4 \left(\sum t\boldsymbol{z}_{t-1}\right)\left(\sum t\boldsymbol{z}_{t-1}'\right),\end{aligned}$$

whereupon (18) may be written as

$$\begin{aligned}\boldsymbol{R} &= \left[\begin{array}{cc} p_1 - d_1 & (\boldsymbol{p}_2 - \boldsymbol{d}_2)' \\ (\boldsymbol{p}_3 - \boldsymbol{d}_3) & \boldsymbol{P}_4 - \boldsymbol{D}_4 \end{array}\right]\left[\begin{array}{cc} q_1 - e_1 & (\boldsymbol{q}_2 - \boldsymbol{e}_2)' \\ (\boldsymbol{q}_3 - \boldsymbol{e}_3) & \boldsymbol{Q}_4 - \boldsymbol{E}_4 \end{array}\right]^{-1} \\ &\equiv \boldsymbol{F}\,\boldsymbol{G}^{-1} = \left[\begin{array}{cc} f_1 & \boldsymbol{f}_2' \\ \boldsymbol{f}_3 & \boldsymbol{F}_4 \end{array}\right]\left[\begin{array}{cc} g^{(1)} & \boldsymbol{g}^{(2)\prime} \\ \boldsymbol{g}^{(3)} & \boldsymbol{G}^{(4)} \end{array}\right] \\ &= \left[\begin{array}{cc} f_1 g^{(1)} + \boldsymbol{f}_2'\boldsymbol{g}^{(3)} & f_1^{(2)\prime}\boldsymbol{g} + \boldsymbol{f}_2'\boldsymbol{G}^{(4)} \\ \boldsymbol{f}_3^{(1)}\boldsymbol{g} + \boldsymbol{F}_4'\boldsymbol{g}^{(3)} & \boldsymbol{f}_3^{(2)\prime}\boldsymbol{g} + \boldsymbol{F}_4'\boldsymbol{G}^{(4)} \end{array}\right].\end{aligned}$$

Now, define $\{\overline{e}_t\}$ to have element $\{-e_{1t}\}$ and $\{e_{jt}\}$, $j = 2, \ldots, k$, whereupon $\{\overline{\boldsymbol{z}}_t\}$ is equal to $\{\boldsymbol{z}_t\}$, except for the first element; and we note that $\boldsymbol{W}$ and $\boldsymbol{W}^{-1}$ do not change under this operation. Clearly,

$$\begin{aligned}\overline{\boldsymbol{P}} &= \left[\begin{array}{cc} p_1 & -\boldsymbol{p}_2' \\ -\boldsymbol{p}_3 & \boldsymbol{P}_4 \end{array}\right], \quad \overline{\boldsymbol{Q}} = \left[\begin{array}{cc} q_1 & -\boldsymbol{q}_2' \\ -\boldsymbol{q}_3 & \boldsymbol{Q}_4 \end{array}\right], \quad \overline{\boldsymbol{D}} = \left[\begin{array}{cc} d_1 & -\boldsymbol{d}_2' \\ -\boldsymbol{d}_3 & \boldsymbol{D}_4 \end{array}\right], \\ \overline{\boldsymbol{E}} &= \left[\begin{array}{cc} e_1 & -\boldsymbol{e}_2' \\ -\boldsymbol{e}_3 & \boldsymbol{E}_4 \end{array}\right], \quad \overline{\boldsymbol{F}} = \left[\begin{array}{cc} f_1 & -\boldsymbol{f}_2' \\ -\boldsymbol{f}_3 & \boldsymbol{F}_4 \end{array}\right], \quad \overline{\boldsymbol{G}} = \left[\begin{array}{cc} g_1 & -\boldsymbol{g}_2' \\ -\boldsymbol{g}_3 & \boldsymbol{G}_4 \end{array}\right],\end{aligned}$$

from which

$$\overline{\boldsymbol{R}} = \left[\begin{array}{cc} f_1 g^{(1)} + \boldsymbol{f}_2'\boldsymbol{g}^{(3)} & -f_1\boldsymbol{g}^{(2)\prime} - \boldsymbol{f}_2'\boldsymbol{G}^{(4)} \\ -\boldsymbol{f}_3 g^{(1)} - \boldsymbol{F}_4\boldsymbol{g}^{(3)} & \boldsymbol{f}_3\boldsymbol{g}^{(2)\prime} + \boldsymbol{F}_4'\boldsymbol{G}^{(4)} \end{array}\right].$$

However, we know that $\boldsymbol{L}\,\mathrm{E}\,(\boldsymbol{R})\,\boldsymbol{L}^{-1} = \boldsymbol{L}\,\mathrm{E}\,\left(\overline{\boldsymbol{R}}\right)\boldsymbol{L}^{-1}$; since corresponding off-diagonal blocks have opposite signs in $\boldsymbol{R}$ and $\overline{\boldsymbol{R}}$, these must also have zero expectations. Upon application of the same logic to all off-diagonal elements, we prove that $\boldsymbol{\Phi}_{\mu,\delta}$ is a diagonal matrix; the diagonal elements are then equal because of the invariance of the normalized VAR specification to the re-ordering of the components of $\boldsymbol{z}_t$; also, bias is not a function of $\boldsymbol{\Omega}$.

References

Abadir, K. M., 1993a. OLS bias in a nonstationary autoregression, Econometric Theory 9, 81-93.

Abadir, K. M., 1993b. Expansions for some confluent hypergeometric functions, Journal of Physics A 26, 4059-4066 (corrigendum for printing error, 1993, 26: 7663).

Abadir, K. M., 1995a. Unbiased estimation as a solution to testing for random walks, Economics Letters 47, 263-268.

Abadir, K. M., 1995b. On efficient simulations in dynamic models, Discussion Paper in Economics 95/21, University of Exeter.

Abadir, K. M., 1999. An introduction to hypergeometric functions for economists, Econometric Reviews 18, 287-330.

Abadir, K. M., Hadri, K., 2000. Is more information a good thing? Bias nonmonotonicity in stochastic difference equations, Bulletin of Economic Research 52, 91-100.

Abadir, K. M., Hadri, K., Tzavalis, E., 1999. The influence of VAR dimensions on estimator biases, Econometrica 67, 163-181.

Abadir, K. M., Larsson, R., 1996. The joint moment generating function of quadratic forms in multivariate autoregressive series, Econometric Theory 12, 682-704.

Abadir, K. M., Larsson, R., 2001. The joint moment generating function of quadratic forms in multivariate autoregressive series: The case with deterministic components, Econometric Theory 17, 222-246.

Abadir, K. M., Magnus, J. R., 2002. Notation in econometrics: A proposal for a standard, Econometrics Journal 5, 76-90.

Andrews, D. W. K., 1993. Exactly median-unbiased estimation of first order autoregressive/unit root models, Econometrica 61, 139-165.

Backus, D., 1986. The Canadian-U.S. exchange rate: Evidence from a vector autoregression, Review of Economics and Statistics 68, 628-637.

Blanchard, O. J., Quah, D., 1989. The dynamic effects of aggregate demand and supply disturbances, American Economic Review 79, 655-673.

Campos, J., 1986. Finite-sample properties of the instrumental-variables estimator for dynamic simultaneous-equation subsystems with ARMA disturbances, Journal of Econometrics 32, 333-366.

Canova, F., 1995. Vector autoregressive models: Specification, estimation, inference, and forecasting, in: Pesaran, M. H., Wickens, M. R. eds., Handbook of Applied Econometrics, Volume 1: Macroeconomics (Blackwell Publishers, Oxford) 73-138.

Davidson, R., MacKinnon, J. G., 1993. Estimation and Inference in Econometrics (Oxford University Press, Oxford).

Dickey, D. A., Fuller, W. A., 1979. Distribution of the estimators for autoregressive time series with a unit root, Journal of the American Statistical Association 74, 427-431.

Dickey, D. A., Fuller, W. A., 1981. Likelihood ratio statistics for autoregressive time series with a unit root, Econometrica 49, 1057-1072.

Engle, R. F., Hendry, D. F., Trumble, D., 1985. Small-sample properties of ARCH estimators and tests, Canadian Journal of Economics 18, 66-93.

Ericsson, N. R., 1991. Monte Carlo methodology and the finite sample properties of instrumental variables statistics for testing nested and non-nested hypotheses, Econometrica 59, 1249-1277.

Ericsson, N. R., MacKinnon, J. G., 2001. Distributions of error correction tests for cointegration, Econometrics Journal 4, 1-34.

Evans, G. B. A., Savin, N. E., 1981. Testing for unit roots: 1, Econometrica 49, 753-779.

Evans, G. B. A., Savin, N. E., 1984. Testing for unit roots: 2, Econometrica 52, 1241-1269.

Fang, K.-T., Kotz, S., Ng, K.-W., 1990. Symmetric Multivariate and Related Distributions (Chapman and Hall, London).

Farebrother, R. W., 1980. The Durbin-Watson test for serial correlation when there is no intercept in the regression, Econometrica 48, 1553-1563.

Forchini, G., 2001. The exact cumulative distribution function of a ratio of quadratic forms in normal variables with application to the AR(1) model, Discussion Paper in Economics, No. 2001/02, University of York.

Gonzalo, J., Pitarakis, J.-Y., 1998. On the exact moments of asymptotic distributions in an unstable AR(1) with dependent errors, International Economic Review 39, 71-88.

Hendry, D. F., 1982. A reply to Professors Maasoumi and Phillips, Journal of Econometrics 19, 203-213.

Hendry, D. F., 1984. Monte Carlo experimentation in econometrics, in: Griliches, Z., Intriligator, M. D. eds., Handbook of Econometrics, Volume 2 (North-Holland, Amsterdam) 937-976.

Hendry, D. F., Trivedi, P. K., 1972. Maximum likelihood estimation of difference equations with moving average errors: A simulation study, Review of Economic Studies 39, 117-145.

Ip, W. C., Phillips, G. D. A., 1998. The non-monotonicity of the bias and mean squared error of the two stage least squares estimators of exogenous variable coefficients, Economics Letters 60, 303-310.

Jarque, C. M., Bera, A. K., 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals, Economics Letters 6, 255-259.

Johansen, S., 1995. Likelihood-based Inference in Cointegrated Vector Autoregressive Models (Oxford University Press, Oxford).

Lawford, S., 2001. Improved modelling in finite-sample and nonlinear frameworks, unpublished D.Phil. thesis, University of York.

Le Breton, A., Pham, D. T., 1989. On the bias of the least squares estimator for the first order autoregressive process, Annals of the Institute of Statistical Mathematics 41, 555-563.

Lütkepohl, H., 1993. Introduction to Multiple Time Series Analysis (Springer-Verlag, Berlin) 2nd edition.

Maasoumi, E., Phillips, P. C. B., 1982. On the behavior of inconsistent instrumental variable estimators, Journal of Econometrics 19, 183-201.

MacKinnon, J. G., 1996. Numerical distribution functions for unit root and cointegration tests, Journal of Applied Econometrics 11, 601-618.

MacKinnon, J. G., 2000. Computing numerical distribution functions in econometrics, in: Pollard, A., Mewhort, D., Weaver, D. eds., High Performance Computing Systems and Applications (Kluwer, Amsterdam) 455-470.

MacKinnon, J. G., Haug, A. A., Michelis, L., 1999. Numerical distribution functions of likelihood ratio tests for cointegration, Journal of Applied Econometrics 14, 563-577.

MacKinnon, J. G., Smith, A. A., Jr., 1998. Approximate bias correction in econometrics, Journal of Econometrics 85, 205-230.

Maeshiro, A., 1999. A lagged dependent variable, autocorrelated disturbances, and unit root tests – peculiar OLS bias properties – a pedagogical note, Applied Economics 31, 381-396.

Magnus, J. R., Neudecker, H., 1999. Matrix Differential Calculus with Applications in Statistics and Econometrics (John Wiley and Sons, Chichester) revised edition.

Mikhail, W. M., 1972. Simulating the small-sample properties of econometric estimators, Journal of the American Statistical Association 67, 620-624.

Mikhail, W. M., 1975. A comparative Monte Carlo study of the properties of econometric estimators, Journal of the American Statistical Association 70, 94-104.

Nankervis, J. C., Savin, N. E., 1988a. The Student's t approximation in a stationary first order autoregressive model, Econometrica 56, 119-145.

Nankervis, J. C., Savin, N. E., 1988b. The exact moments of the least squares estimator for the autoregressive model: Corrections and extensions, Journal of Econometrics 37, 381-388.

Ooms, M., 1994. Empirical Vector Autoregressive Modelling (Springer-Verlag, Berlin) Lecture Notes in Economics and Mathematical Systems, No. 407.

Park, J. Y., Phillips, P. C. B., 1988. Statistical inference in regressions with integrated processes: Part 1, Econometric Theory 4, 468-497.

Park, J. Y., Phillips, P. C. B., 1989. Statistical inference in regressions with integrated processes: Part 2, Econometric Theory 5, 95-131.

Patterson, K. D., 2000. Bias reduction in autoregressive models, Economics Letters 68, 135-141.

Pere, P., 2000. Adjusted estimates and Wald statistics for the AR(1) model with constant, Journal of Econometrics 98, 335-363.

Phillips, P. C. B., 1980. Finite sample theory and the distributions of alternative estimators of the marginal propensity to consume, Review of Economic Studies 47, 183-224.

Phillips, P. C. B., 1987a. Asymptotic expansions in nonstationary vector autoregressions, Econometric Theory 3, 45-68.

Phillips, P. C. B., 1987b. Time series regression with a unit root, Econometrica 55, 277-301.

Phillips, P. C. B., 1995. Nonstationary time series and cointegration, Journal of Applied Econometrics 10, 87-94.

Phillips, P. C. B., Durlauf, S. N., 1986. Multiple time series regression with integrated processes, Review of Economic Studies 53, 473-495.

Sims, C. A., 1980. Macroeconomics and reality, Econometrica 48, 1-48.

Stamatogiannis, M. P., 1999. A study on the finite sample bias of an unstable VAR with a drift, unpublished M.Sc. thesis, University of York.

Tanizaki, H., 2000. Bias correction of OLSE in the regression model with lagged dependent variables, Computational Statistics and Data Analysis 34, 495-511.

Tsay, R. S., Tiao, G. C., 1990. Asymptotic properties of multivariate nonstationary processes with applications to autoregressions, Annals of Statistics 18, 220-250.

Tsui, A. K., Ali, M. M., 1989. Exact moments of the least squares estimator in a first-order non-stationary autoregressive model, Proceedings of the American Statistical Association, Business and Economics Section, 220-225.

Tsui, A. K., Ali, M. M., 1994. Exact distributions, density functions and moments of the least-squares estimator in a first-order autoregressive model, Computational Statistics and Data Analysis 17, 433-454.

Urzúa, C. M., 1996. On the correct use of omnibus tests for normality, Economics Letters 53, 247-251 (corrigendum, 1997, 54: 301).

Vinod, H. D., Shenton, L. R., 1996. Exact moments for autoregressive and random walk models for a zero or stationary initial value, Econometric Theory 12, 481-499.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica 48, 817-838.

White, J. S., 1958. The limiting distribution of the serial correlation coefficient in the explosive case, Annals of Mathematical Statistics 29, 1188-1197.

Table 1: Bias in Case A: $-100 \times \widetilde{b}$ $\left(-100 \times b^{AHT}\right)$ $\left[-100 \times b^{RS}\right]$

T	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
25	6.4	13.5	20.0	26.1	31.8	37.1	42.1	46.7
	(6.4)	(12.8)	(19.3)	(25.7)	(32.1)	(38.5)	(44.9)	(51.3)
	[6.5]	[13.5]	[20.0]	[26.2]	[31.9]	[37.3]	[42.3]	[46.9]
50	3.4	7.2	10.8	14.3	17.6	20.9	24.0	27.0
	(3.4)	(6.8)	(10.1)	(13.5)	(16.9)	(20.3)	(23.7)	(27.1)
	[3.4]	[7.1]	[10.8]	[14.3]	[17.8]	[21.1]	[24.3]	[27.4]
100	1.7	3.7	5.6	7.5	9.3	11.1	12.9	14.6
	(1.7)	(3.5)	(5.2)	(6.9)	(8.7)	(10.4)	(12.1)	(13.9)
	[1.7]	[3.7]	[5.6]	[7.5]	[9.4]	[11.2]	[13.0]	[14.8]
200	0.9	1.9	2.9	3.8	4.8	5.8	6.7	7.6
	(0.9)	(1.8)	(2.6)	(3.5)	(4.4)	(5.3)	(6.2)	(7.0)
	[0.9]	[1.9]	[2.9]	[3.8]	[4.8]	[5.8]	[6.7]	[7.7]
400	0.4	0.9	1.4	1.9	2.4	2.9	3.4	3.9
	(0.4)	(0.9)	(1.3)	(1.8)	(2.2)	(2.7)	(3.1)	(3.5)
	[0.4]	[0.9]	[1.4]	[1.9]	[2.4]	[2.9]	[3.4]	[3.9]
800	0.2	0.5	0.7	1.0	1.2	1.5	1.7	2.0
	(0.2)	(0.4)	(0.7)	(0.9)	(1.1)	(1.3)	(1.6)	(1.8)
	[0.2]	[0.5]	[0.7]	[1.0]	[1.2]	[1.5]	[1.7]	[2.0]

Table 2: Bias in Cases A,B,C: $-100 \times \widetilde{b}$ $\left(-100 \times \widetilde{b}_{\mu}\right)$ $\left[-100 \times \widetilde{b}_{\mu,\delta}\right]$

T	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
25	6.4	13.5	20.0	26.1	31.8	37.1	42.1	46.7
	(19.2)	(25.0)	(30.6)	(35.9)	(40.9)	(45.7)	(50.2)	(54.5)
	[35.3]	[40.0]	[44.5]	[49.0]	[53.2]	[57.3]	[61.2]	[64.9]
50	3.4	7.2	10.8	14.3	17.6	20.9	24.0	27.0
	(10.1)	(13.4)	(16.7)	(19.9)	(23.0)	(26.0)	(28.9)	(31.8)
	[19.0]	[21.8]	[24.7]	[27.5]	[30.3]	[33.0]	[35.7]	[38.3]
100	1.7	3.7	5.6	7.5	9.3	11.1	12.9	14.6
	(5.2)	(7.0)	(8.7)	(10.5)	(12.2)	(14.0)	(15.7)	(17.3)
	[9.9]	[11.4]	[13.0]	[14.6]	[16.3]	[17.9]	[19.5]	[21.1]
200	0.9	1.9	2.9	3.8	4.8	5.8	6.7	7.6
	(2.6)	(3.6)	(4.5)	(5.4)	(6.3)	(7.3)	(8.2)	(9.1)
	[5.0]	[5.8]	[6.7]	[7.6]	[8.4]	[9.3]	[10.2]	[11.1]
400	0.4	0.9	1.4	1.9	2.4	2.9	3.4	3.9
	(1.3)	(1.8)	(2.3)	(2.7)	(3.2)	(3.7)	(4.2)	(4.6)
	[2.5]	[3.0]	[3.4]	[3.9]	[4.3]	[4.8]	[5.2]	[5.7]
800	0.2	0.5	0.7	1.0	1.2	1.5	1.7	2.0
	(0.7)	(0.9)	(1.1)	(1.4)	(1.6)	(1.9)	(2.1)	(2.4)
	[1.3]	[1.5]	[1.7]	[1.9]	[2.2]	[2.4]	[2.6]	[2.9]

Table 3: Bias $-100 \times \widetilde{b}$ in Case A ($\boldsymbol{x}_0 \neq \mathbf{0}_k$)

T	k	$\boldsymbol{x}_0'\boldsymbol{x}_0 = 10$	$\boldsymbol{x}_0'\boldsymbol{x}_0 = 20$	$\boldsymbol{x}_0'\boldsymbol{x}_0 = 50$	$\boldsymbol{x}_0'\boldsymbol{x}_0 = 100$	$\boldsymbol{x}_0'\boldsymbol{x}_0 = 150$
25	2	11.8	11.0	10.2	9.9	9.8
	3	18.4	17.8	17.1	16.9	16.8
	4	24.5	23.9	23.4	23.2	23.1
	5	30.2	29.7	29.2	28.9	28.9
	6	35.5	35.0	34.5	34.3	34.2
50	2	6.6	6.2	5.7	5.4	5.3
	3	10.2	9.9	9.5	9.2	9.1
	4	13.8	13.4	13.0	12.8	12.7
	5	17.1	16.8	16.4	16.2	16.1
	6	20.3	20.0	19.6	19.4	19.3
100	2	3.5	3.4	3.2	3.0	2.9
	3	5.4	5.3	5.1	4.9	4.8
	4	7.3	7.2	7.0	6.8	6.8
	5	9.2	9.0	8.8	8.7	8.6
	6	11.0	10.8	10.6	10.5	10.4
200	2	1.8	1.8	1.7	1.6	1.5
	3	2.8	2.8	2.7	2.6	2.5
	4	3.8	3.8	3.7	3.6	3.5
	5	4.8	4.7	4.6	4.6	4.5
	6	5.7	5.7	5.6	5.5	5.5

Table 4: Variance in Cases A,B,C: $10000 \times \widetilde{V}$ $\left(10000 \times \widetilde{V}_{\mu}\right)$ $\left[10000 \times \widetilde{V}_{\mu,\delta}\right]$

T	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
25	132.4	251.0	351.5	440.1	519.6	593.0	663.0	730.7
	(228.5)	(318.2)	(398.0)	(470.4)	(537.0)	(599.4)	(660.1)	(719.9)
	[339.6]	[417.4]	[486.2]	[549.9]	[610.9]	[668.5]	[727.0]	[786.6]
50	36.5	70.9	102.3	130.4	156.8	181.0	203.6	225.0
	(67.5)	(96.4)	(123.5)	(148.5)	(171.9)	(193.7)	(214.0)	(233.3)
	[109.8]	[135.8]	[160.0]	[183.0]	[204.5]	[224.4]	[242.9]	[260.5]
100	9.6	18.8	27.8	36.2	44.2	51.8	59.1	66.2
	(18.5)	(26.7)	(35.0)	(42.7)	(50.2)	(57.3)	(64.3)	(71.0)
	[31.4]	[39.4]	[47.3]	[54.6]	[61.7]	[68.5]	[75.0]	[81.4]
200	2.5	4.9	7.3	9.6	11.8	14.0	16.1	18.2
	(4.8)	(7.1)	(9.3)	(11.5)	(13.7)	(15.8)	(17.8)	(19.9)
	[8.5]	[10.7]	[12.9]	[15.0]	[17.1]	[19.2]	[21.2]	[23.2]
400	0.6	1.2	1.9	2.5	3.1	3.6	4.2	4.8
	(1.2)	(1.8)	(2.4)	(3.0)	(3.6)	(4.2)	(4.7)	(5.3)
	[2.2]	[2.8]	[3.4]	[4.0]	[4.5]	[5.1]	[5.6]	[6.2]
800	0.2	0.3	0.5	0.6	0.8	0.9	1.1	1.2
	(0.3)	(0.5)	(0.6)	(0.8)	(0.9)	(1.1)	(1.2)	(1.4)
	[0.6]	[0.7]	[0.9]	[1.0]	[1.2]	[1.3]	[1.5]	[1.6]

Table 5: Correction factors λ for $\widehat{\boldsymbol{A}}$ that yield minimum MSE

T	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
25	1.05	1.12	1.19	1.25	1.32	1.38	1.44	1.49
50	1.03	1.07	1.11	1.15	1.19	1.23	1.27	1.31
100	1.02	1.04	1.06	1.08	1.10	1.12	1.14	1.16
200	1.01	1.02	1.03	1.04	1.05	1.06	1.07	1.08
400	1.00	1.01	1.01	1.02	1.02	1.03	1.03	1.04
800	1.00	1.00	1.01	1.01	1.01	1.01	1.02	1.02

Table 6: Corrected $-100 \times b^m$ in Case A; $\frac{b^m}{\widetilde{b}}$ in $[\cdot]$

T	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
25	1.49	3.24	5.21	7.46	10.05	13.04	16.50	20.48
	[0.23]	[0.24]	[0.26]	[0.29]	[0.32]	[0.35]	[0.39]	[0.44]
50	0.39	0.82	1.27	1.74	2.26	2.81	3.40	4.05
	[0.12]	[0.11]	[0.12]	[0.12]	[0.13]	[0.13]	[0.14]	[0.15]
100	0.10	0.20	0.31	0.42	0.53	0.65	0.77	0.90
	[0.06]	[0.05]	[0.06]	[0.06]	[0.06]	[0.06]	[0.06]	[0.06]
200	0.03	0.05	0.08	0.10	0.13	0.16	0.18	0.21
	[0.03]	[0.03]	[0.03]	[0.03]	[0.03]	[0.03]	[0.03]	[0.03]
400	0.01	0.01	0.02	0.03	0.03	0.04	0.05	0.05
	[0.01]	[0.01]	[0.01]	[0.01]	[0.01]	[0.01]	[0.01]	[0.01]
800	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01
	[0.01]	[0.01]	[0.01]	[0.01]	[0.01]	[0.01]	[0.01]	[0.01]

Table 7: Corrected $10,000 \times V^m$ in Case A; $\frac{V^m}{\widetilde{V}}$ in $[\cdot]$

T	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
25	139.4	280.6	416.6	551.4	685.2	819.9	955.6	1090.9
	[1.05]	[1.12]	[1.19]	[1.25]	[1.32]	[1.38]	[1.44]	[1.49]
50	37.6	75.7	113.2	149.5	186.1	222.3	258.7	295.7
	[1.03]	[1.07]	[1.11]	[1.15]	[1.19]	[1.23]	[1.27]	[1.31]
100	9.8	19.5	29.4	38.9	48.5	57.9	67.3	76.9
	[1.02]	[1.04]	[1.06]	[1.08]	[1.10]	[1.12]	[1.14]	[1.16]
200	2.5	5.0	7.5	10.0	12.4	14.8	17.2	19.7
	[1.01]	[1.02]	[1.03]	[1.04]	[1.05]	[1.06]	[1.07]	[1.08]
400	0.6	1.3	1.9	2.5	3.1	3.8	4.4	5.0
	[1.00]	[1.01]	[1.01]	[1.02]	[1.02]	[1.03]	[1.03]	[1.04]
800	0.2	0.3	0.5	0.6	0.8	0.9	1.1	1.3
	[1.00]	[1.00]	[1.01]	[1.01]	[1.01]	[1.01]	[1.02]	[1.02]

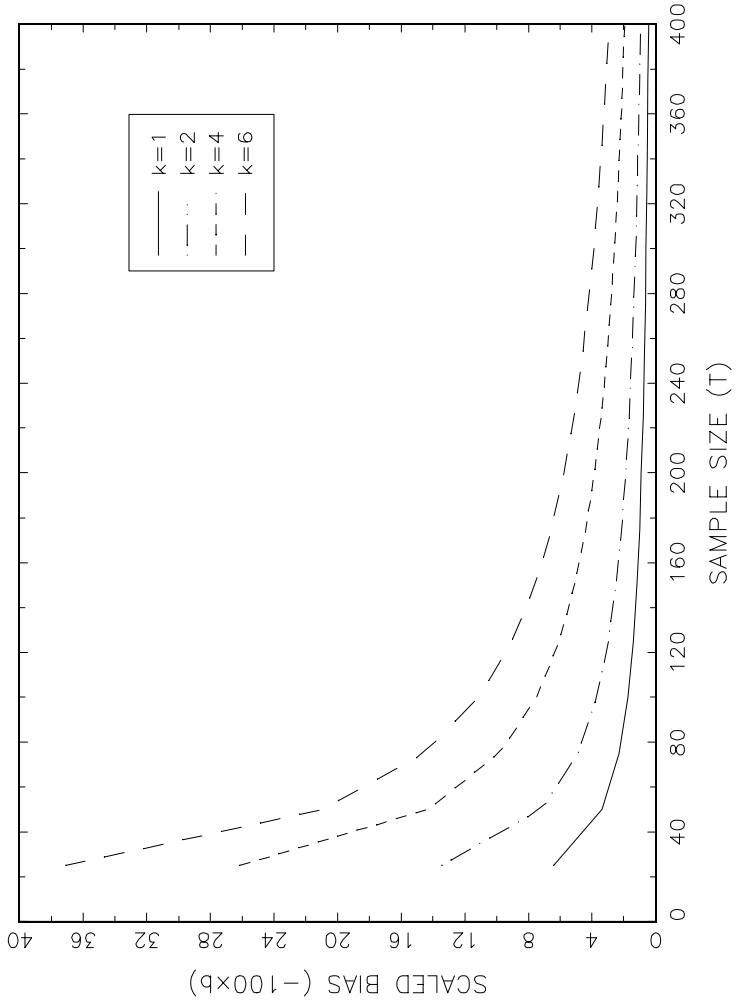

Figure 1: $-100 \times \widetilde{b}$ against T, in Case A

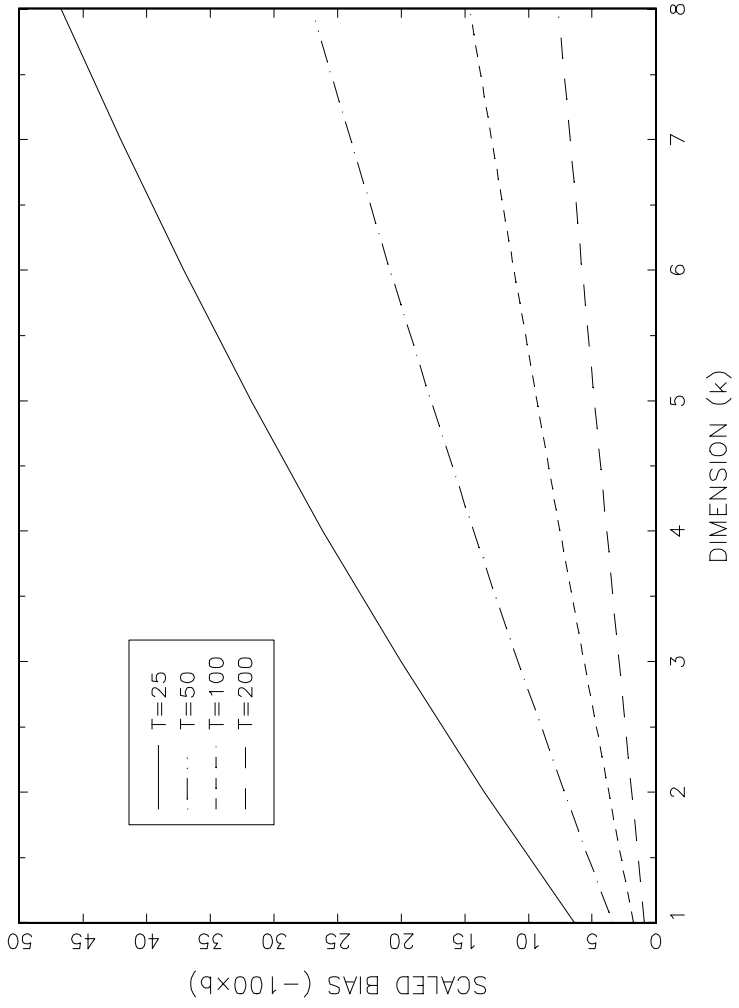

Figure 2: $-100 \times \widetilde{b}$ against k, in Case A

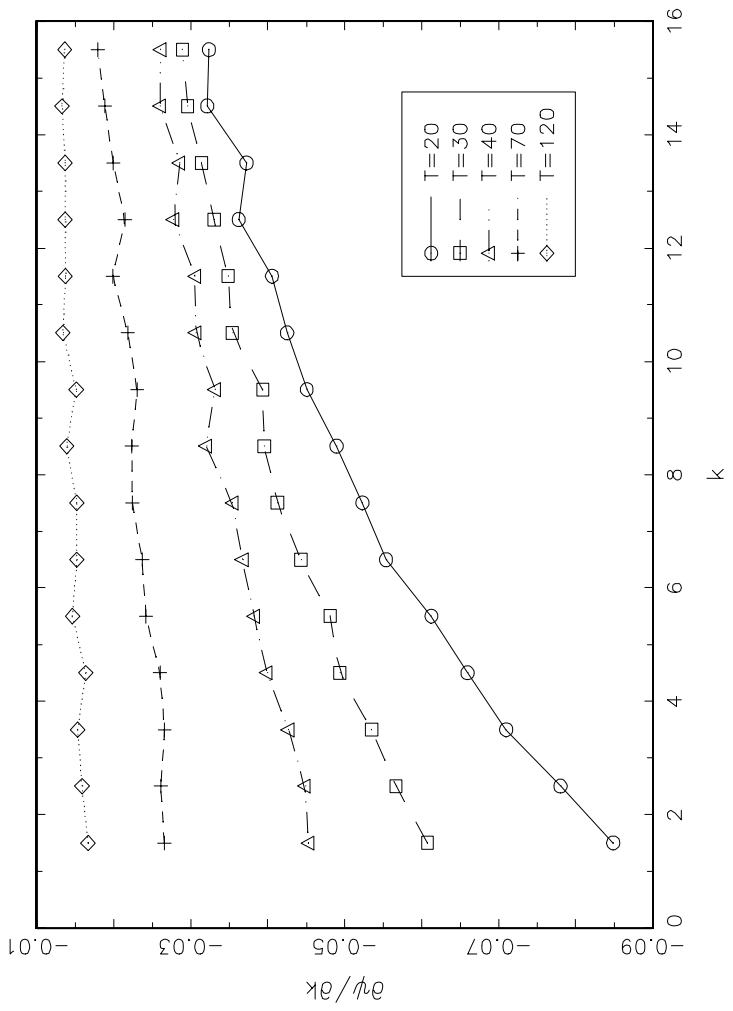

Figure 3: $\partial\widetilde{\psi}/\partial k$ against k in Case A

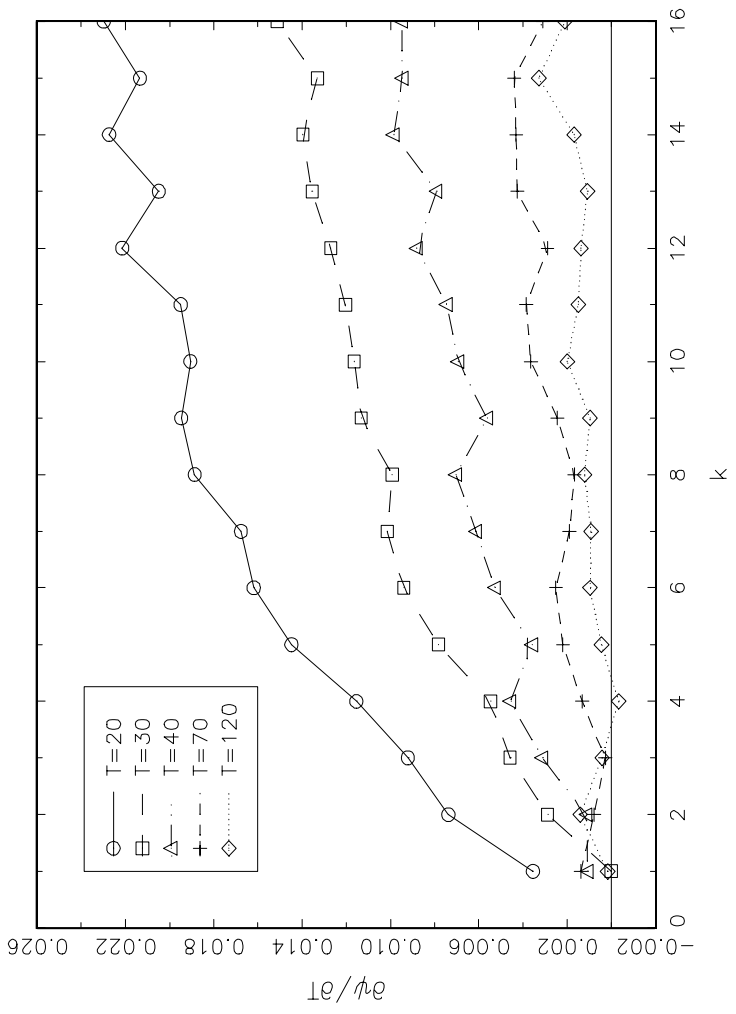

Figure 4: $\partial\widetilde{\psi}/\partial T$ against k in Case A

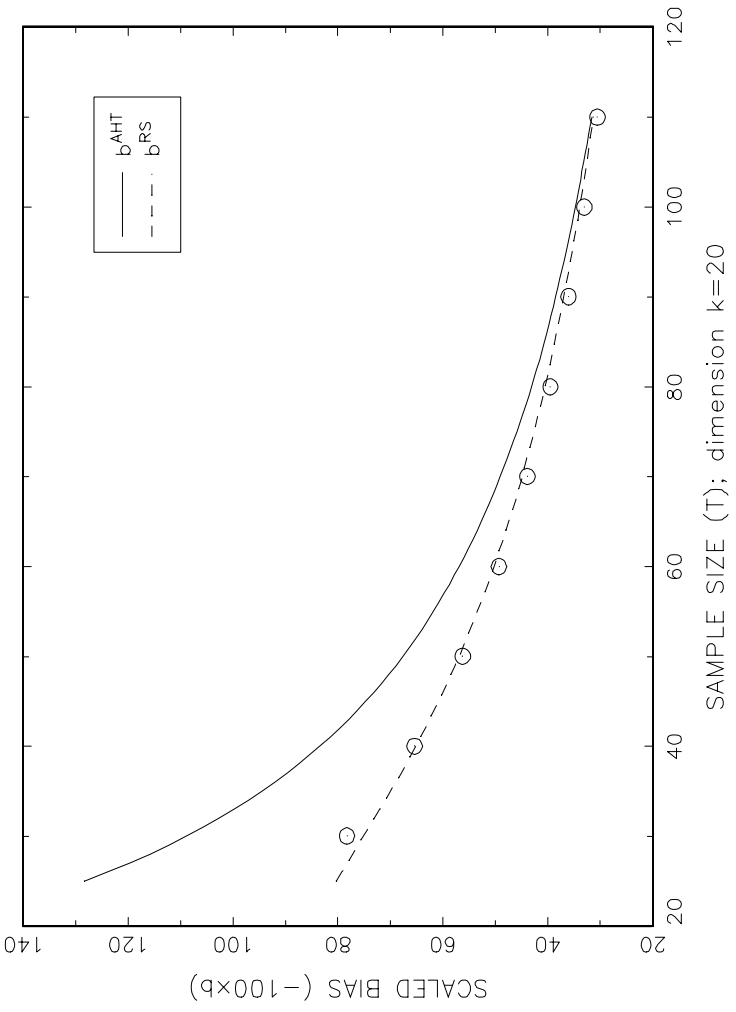

Figure 5: b^{AHT} and b^{RS}; out-of-sample fit ($k = 20$)

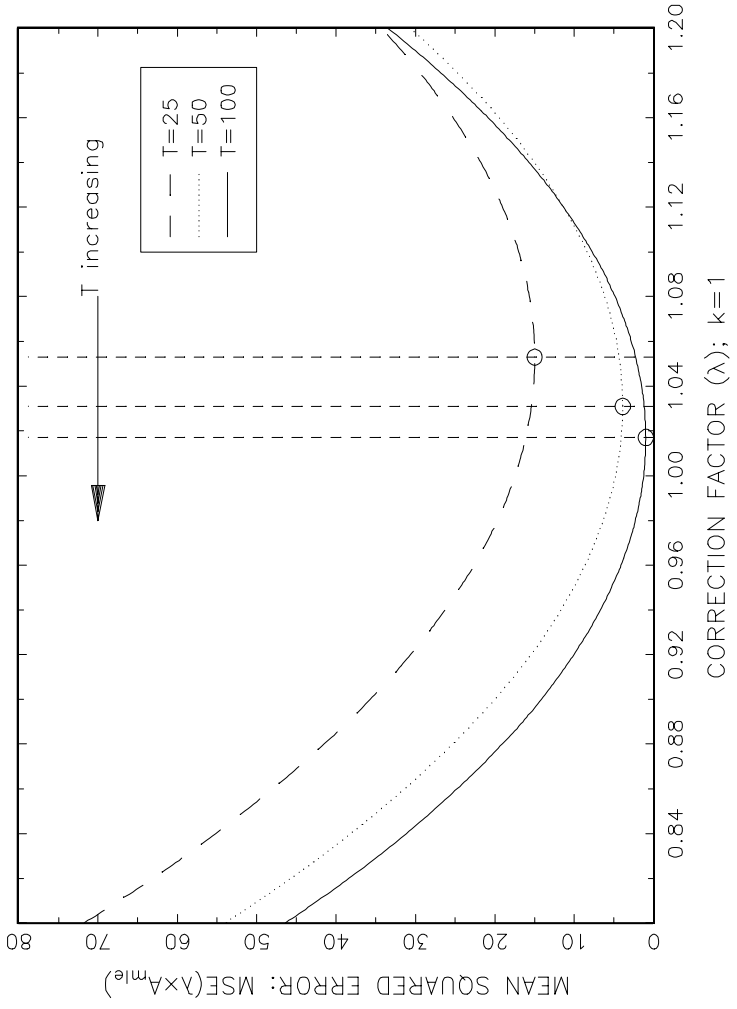

Figure 6: Mean squared error, various λ ($k = 1$)